Silvano Salvador

Toponomastic Networks

Closing the Whole-Country Gap: Delimited Substrate Zones from Real GeoNames Corpora

Volume III

A Companion Series to Toponomastic Entropy

Contents

Preface **vi**

1 A Verifiable Delimitation Method **1**
1.1 The Admin1-Lookup Problem 1
1.2 The Delimitation Protocol 1

2 Central Anatolia (`TR.txt`) **3**
2.1 Zone Definition 3
2.2 Pool Composition and Aggregate Statistics 3
2.3 Roadmap: Further Zones Reachable from Data Already at Hand . . . 5

3 Narrowing to Cappadocia **7**
3.1 Zone Definition and Motivation 7
3.2 Pool Composition and Aggregate Statistics 7
3.3 Discussion: Explanation (b) Is Supported 8
3.4 Feature-Class Decomposition: Settling the Open Question 9

4 The Alpine German-Speaking Core: Switzerland and Austria **11**
4.1 A Verified Linguistic Delimitation 11
4.2 Pool Composition and Comparison to the Whole-Country Baseline . 12
4.3 Cross-Reference to Chapter 5 13

5 Elevation as a Data-Native Delimitation Criterion **14**
5.1 Motivation: A Delimitation Requiring No Admin Lookup at All . . . 14
5.2 Coverage Caveat 14
5.3 The Elevation Split 15
5.4 Elevation Versus Linguistic Delimitation: A Direct Comparison . . . 15

6 The British Isles and the Celtic Fringe **17**
6.1 A Delimitation Requiring No Interpretation at All 17
6.2 Six Pools, One Country Group 17

7 Ireland's Four Historic Provinces **20**
7.1 Four Provinces, Self-Verified 20
7.2 Pool Composition and Aggregate Statistics 21
7.3 Roadmap 22
7.4 Reproducibility: What Was Actually Run 22

8 Reunified Ulster Comparison **23**
8.1 A Nine-County Pool Spanning Two Country Files 23
8.2 A Direct Cross-Border JSD 24
8.3 Reproducibility: What Was Actually Run 24

9 A Maghreb Coastal-Versus-Interior Split **25**
9.1 Extending the Protocol to Three Further Countries 25
9.2 Results 26
9.3 Does the Whole-Country Maghreb Clustering Survive the Split? . . . 27

10 The Caucasus: Reproducibility, Elevation, Clustering **28**
10.1 Motivation: A Region With No Internal Admin1 Split to Verify 28
10.2 Reproducibility Check: Recomputing Three Whole-Country Pools . . 29
10.3 Coverage Caveat: Elevation Reporting Is Far Sparser Here Than in the Alps 29
10.4 The Elevation Split, Two Ways 30
10.5 Cross-National Clustering: Do the Three Countries Cohere as a Group? 31
10.6 Roadmap Addendum 32
10.7 Reproducibility: What Was Actually Run 32

11 Caucasus Roadmap: Sub-National Zones and Elevation Sweep **33**
11.1 Two Declared Sub-National Delimitations 33
11.2 Pool Composition and Results 34
11.3 The Elevation-Threshold Sweep 35
11.4 What Remains Open 36

12 Roadmap: Resampling, Elevation Redraw, Armenian Zone **38**
12.1 A Resampling Check of Talysh's H_1 Lifetime-to-Scale Ratio 38
12.2 An Independent-Redraw Variance Estimate for the Elevation Sweep's Two Most Extreme Thresholds 40
12.3 Toward an Armenian Delimitation: A Verified Yazidi Settlement List, Not Yet a Complete Zone 41
12.4 Updated Roadmap 42
12.5 Reproducibility: What Was Actually Run 43

13 Armenian ADM2 Zone and Talysh Resample **44**
13.1 Completing the Armenian Delimitation: From Twenty-Two Villages to a Six-Community ADM2 Pool 44
13.2 A Fifty-Draw Extension of the Talysh Resampling Check 46
13.3 Reproducibility: What Was Actually Run 47

14 The Iberian Peninsula's Autonomous Communities **49**
14.1 A Delimitation With No Judgement Call 49
14.2 Pool Composition and Whole-Country Baselines 50
14.3 Cross-Regional Distances: A Genuine Galician–Portuguese Signal . . 51
14.4 A Population-Level Cross-Check: Jensen–Shannon Divergence 52
14.5 Roadmap 53
14.6 Reproducibility: What Was Actually Run 53

15 The Western Balkans: Istria and Dalmatia **55**
15.1 Two Delimitations, Two Different Levels of External Judgement . . . 55
15.2 Pool Composition and Aggregate Statistics . . . 56
15.3 Cross-Border Distances at the Istrian Peninsula . . . 57
15.4 Roadmap . . . 58
15.5 Reproducibility: What Was Actually Run . . . 59

16 Pakistan's Provinces **60**
16.1 Seven Admin1 Units, Self-Verified . . . 60
16.2 Pool Composition and Aggregate Statistics . . . 61
16.3 Cross-Provincial Distances: Sindh as an Outlier . . . 61
16.4 A Population-Level Cross-Check . . . 62
16.5 Roadmap . . . 63
16.6 Reproducibility: What Was Actually Run . . . 63

17 France: Brittany, Corsica, and Occitanie **64**
17.1 Three Regions, Self-Verified . . . 64
17.2 Pool Composition and Aggregate Statistics . . . 65
17.3 Cross-Regional Distances: Corsica Stands Apart . . . 65
17.4 A Population-Level Cross-Check . . . 66
17.5 Roadmap . . . 67
17.6 Reproducibility: What Was Actually Run . . . 67

18 India: Ten States and a Clustering Test **68**
18.1 Ten States, Self-Verified . . . 68
18.2 Pool Composition and Aggregate Statistics . . . 69
18.3 A Three-Way Clustering Test . . . 70
18.4 Classical MDS: A Two-Dimensional Projection . . . 71
18.5 A Population-Level Cross-Check . . . 72
18.6 Roadmap . . . 72
18.7 Reproducibility: What Was Actually Run . . . 72

19 Italy: Five Regions, Including the Author's Own **74**
19.1 Five Regions, Self-Verified . . . 74
19.2 Pool Composition and Aggregate Statistics . . . 75
19.3 Cross-Regional Distances: Sardinia's Isolation . . . 75
19.4 A Population-Level Cross-Check . . . 76
19.5 Roadmap . . . 77
19.6 Reproducibility: What Was Actually Run . . . 77

20 Iberia Roadmap: Catalonia Split, Basque Check **78**
20.1 Catalonia's Four Provinces, Recovered at `admin2` Resolution . . . 78
20.2 A Twelve-Draw Resampling Check on Basque's Isolation . . . 79
20.3 Roadmap . . . 81
20.4 Reproducibility: What Was Actually Run . . . 81

21 A Jensen–Shannon Divergence Survey **82**
21.1 The Same Function, Applied Nine Times . . . 82
21.2 Results . . . 83
21.3 What Remains Open . . . 84
21.4 Reproducibility: What Was Actually Run . . . 84

22 JSD Roadmap, Part One: The Maghreb Extension **86**
22.1 Re-Deriving the Coastal Sets, String by String . . . 86
22.2 The Interior Sets: A Second Construction, Not a Second Lookup . . . 87
22.3 The Jensen–Shannon Divergence, Extended to All Three Countries . . . 88
22.4 Reproducibility: What Was Actually Run . . . 88

23 JSD Roadmap, Part Two: The Armenian Outlier **92**
23.1 A Uniform-Random $n = 52$ Null Distribution . . . 92
23.2 What This Resolves, and What It Leaves Open . . . 93
23.3 Reproducibility: What Was Actually Run . . . 94

24 A Volume-Wide Jensen–Shannon Table, Twenty-One Whole-Country Pools **95**
24.1 The One New Computation: India Against Its Own Whole-Country Pool . . . 95
24.2 The Combined Table . . . 96
24.3 Reproducibility: What Was Actually Run . . . 98

25 Libya's Three Historic Provinces **99**
25.1 An Outlier Country, Not Yet Delimited . . . 99
25.2 Zone Definition and Admin1 Verification . . . 100
25.3 Pool Composition and Aggregate Statistics . . . 100
25.4 Cyrenaica's Outlier Ratio . . . 101
25.5 Pairwise Fréchet Distances and Cross-Province Comparison . . . 102
25.6 An External Comparison: Do the Provinces Cluster with the Maghreb? 103
25.7 Roadmap . . . 104
25.8 Reproducibility: What Was Actually Run . . . 104

26 Mauritania: A Saharan/Sahelian Delimitation **105**
26.1 An Omission from the Maghreb Chapter, Revisited . . . 105
26.2 Zone Definition and Admin1 Verification . . . 106
26.3 Pool Composition and Aggregate Statistics . . . 106
26.4 Fréchet Distance and a Maghreb Cross-Check . . . 107
26.5 Roadmap . . . 108
26.6 Reproducibility: What Was Actually Run . . . 108

27 A Named-Strata Permutation Test: Libya and Mauritania **109**
27.1 Method: The Same Test, Named Strata . . . 110
27.2 Results . . . 110
27.3 Reproducibility: What Was Actually Run . . . 111

28 A Complete Twenty-Six-Pool Jensen–Shannon Matrix **113**
28.1 Method: The Same Deterministic Function, Five New Inputs 113
28.2 The Extremes of the Twenty-Six-Pool Matrix 114
28.3 Hierarchical Clustering on the Full Matrix 115
28.4 Reproducibility: What Was Actually Run 117

29 Conclusion: What Twenty-Six Pools Show That One Could Not **119**
29.1 The Preface's Aim, Revisited 119
29.2 A Recurring Pattern: The Same Handful of Pools Keep Isolating Themselves 119
29.3 A Recurring Caution: Two Measures Do Not Always Agree, and Neither Is Simply Right 120
29.4 Elevation: A Genuine but Narrow Data-Native Criterion 120
29.5 What Is Deliberately Left Open 121

A Reproducibility: The Zone-Verification and Delimitation Code **122**

Preface: Why a Third Volume, and Why This Direction

Volumes I and II of *Toponomastic Networks* applied, in turn, graph theory, information geometry and topological data analysis, stochastic processes, and spectral methods to the same underlying GeoNames-derived toponymic material. Volume II's Chapter 9 closed a first data gap by extending real-data validation from three zones to twenty-one whole national GeoNames pools, and went on to show, honestly and with the caveats it stated throughout, that a real and geographically coherent inter-country clustering signal emerges from these pools even though none of them is a delimited substrate zone in *Toponomastic Entropy*'s original sense.

That final qualification is the starting point of this volume. Every real-data chapter so far in this series has worked either with a small number of hand-delimited zones (Basque, Armenian, Libya, and the Basque zone's French-department extension) or with entire national pools, and has flagged, every time whole-country pools were used, that a whole country is not a substrate zone: it is whatever mixture of historical layers its modern political boundary happens to enclose. This volume's stated aim is to close that gap directly, zone by zone, using only real GeoNames data and only delimitation criteria that can be verified from the data files themselves — never from an external lookup table not supplied alongside them, and never invented for convenience.

Methodological Transparency: Scope of This Volume

What follows is no longer an opening test case but a completed twenty-nine-chapter volume, applying the same verifiable delimitation method to zone after zone: the Alpine German-speaking core, the British Isles and Ireland's four historic provinces, a Maghreb coastal/interior split, three Caucasus countries with several sub-national follow-ups, the Iberian Peninsula's co-official-language communities, the Western Balkans, Pakistan, France, India, Italy, Libya, and Mauritania, alongside the two Turkish delimitations that opened the series. Every one of these zones is built, as Chapter 1 sets out, only from codes verified inside the GeoNames files themselves, never from an external lookup table and never invented for convenience.

Repeating the same protocol across so many zones is a methodological choice, not padding: holding the pipeline constant while the input zone changes is what makes each new delimitation a genuine replication rather than a one-off result, and it is also what makes the volume's real payoff possible. That payoff sits in the chapters that step back from individual zones to compare across all of them at

once – the volume-wide Jensen–Shannon surveys, the complete pairwise divergence matrices, and a named-strata permutation test – which only mean anything because every zone feeding them was delimited and verified the same way. The volume closes with a concluding chapter synthesizing what recurred across all twenty-six pools, followed by an appendix laying out the zone-verification and delimitation code in full.

Chapter 1

A Verifiable Delimitation Method

1.1 The Admin1-Lookup Problem

GeoNames' `admin1` field (column 11 of the standard national dump format) is a short alphanumeric code, not a province name. The standard way to interpret it is via GeoNames' own `admin1CodesASCII.txt` lookup table, which was not supplied alongside the national dump files used throughout this series. Guessing the code-to-province correspondence from an external source (for instance, assuming GeoNames' admin1 codes match a country's official postal or vehicle-registration province codes) is not safe: a direct check on `TR.txt`, reported below, shows this assumption is false for Turkey specifically, and there is no general guarantee it holds for any other country either.

Remark: A Verified Counter-Example

Turkey's postal/vehicle plate code for Ankara province is well known to be `06`. A direct query of `TR.txt` for rows with `admin1=06` returns zero rows. The GeoNames admin1 code actually used for Ankara in this file is `68` – verified not by assumption but by locating Ankara's own `ADM1`-feature-class row within `TR.txt` itself, which self-reports its own admin1 code in column 11. This is not a data-quality complaint about GeoNames (its internal admin1 numbering is a legitimate, documented convention, simply a different one from the public plate-code numbering); it is a methodological note explaining why this chapter's delimitation was built exclusively from each province's own self-reported `ADM1` row rather than from any external or assumed code table.

1.2 The Delimitation Protocol

For a target zone defined by a list of province names, the protocol used throughout this volume is:

1. For each province name, locate its own row in the national dump with `feature code = ADM1`; read its self-reported `admin1` code directly from column 11 of that row. This step requires no external file and no assumption about code numbering conventions.

2. Restrict the pool to rows whose `admin1` field matches one of the codes collected in step 1, feature class P or T, exactly as in every other pool-construction step in this series (Volume II, §8.1).

3. Apply the unchanged normalisation, entropy, Fisher–Rao, Fréchet-mean, and persistent-homology pipeline of Volume II, Appendix 8's code (reproduced in condensed form in Appendix A below) to the resulting delimited pool.

This protocol is deliberately conservative: it delimits zones at province (admin1) granularity only, because finer subdivisions (admin2/admin3) would require the same self-verification step repeated recursively and have not yet been checked for reliability in this way. Province-level delimitation is coarser than *Toponomastic Entropy*'s original zone definitions in some cases and comparable in others; where it is coarser, this is stated explicitly rather than implied to be a perfect match.

Chapter 2

Case Study: A Central Anatolian Zone from `TR.txt`

2.1 Zone Definition

Eleven provinces were selected as a plausible interior-plateau delimitation of "Central Anatolia" on ordinary geographic grounds (the classical *İç Anadolu* region: a contiguous plateau block excluding the Aegean, Mediterranean, Black Sea, and southeastern coastal/border provinces): Ankara, Konya, Kayseri, Sivas, Yozgat, Kırşehir, Nevşehir, Niğde, Aksaray, Çankırı, and Karaman. Each province's admin1 code was recovered from its own `ADM1` row in `TR.txt` following the protocol of Section 1.2; the eleven codes are $\{68, 71, 38, 58, 66, 40, 50, 73, 75, 82, 78\}$ respectively, verified individually and not assumed to follow any particular numeric pattern (they manifestly do not: compare Ankara's 68 against Karaman's 78, adjacent alphabetically-sorted province names with no numeric relationship in the code space).

Methodological Transparency: This Is a Fresh Delimitation, Not a Verified Historical Match

The eleven-province list above was assembled for this chapter from general geographic knowledge of the Anatolian plateau's provincial boundaries, not recovered from a specific numbered zone definition in an earlier, unverified volume. No claim is made here that this particular eleven-province set reproduces any specific prior zone boundary; it is offered as a reasonable, geographically motivated, and fully reproducible delimitation in its own right, with the exact province list stated openly so that a different, more principled delimitation (for instance, one derived from a specific published dialectological source) could be substituted in a future revision without altering the pipeline.

2.2 Pool Composition and Aggregate Statistics

Table 2.1 reports the number of pool entries (feature classes P+T, normalised name length ≥ 3) contributed by each of the eleven provinces.

Table 2.2 places the delimited zone's headline statistics alongside the whole-country `TR.txt` pool already computed in Volume II, Table 9.1 (there labelled simply "Turkey").

Table 2.1: Central Anatolia zone: pool contribution by province, sorted by size. Total delimited pool: $9{,}492$ entries.

Province	Entries
Ankara	2,247
Sivas	1,634
Konya	1,410
Yozgat	979
Kayseri	768
Karaman	692
Çankırı	641
Kırşehir	345
Aksaray	291
Nevşehir	253
Niğde	232

Table 2.2: Central Anatolia (delimited, eleven provinces) against the whole-country Turkey pool of Volume II, §9.1.

Quantity	Central Anatolia (delimited)	Turkey (whole country)
Pool n	9,492	70,260
Mean name length	9.81	9.18
Aggregate H_2 (bits)	7.450	7.494
Fisher–Rao mean (n=40)	0.2448	0.2370
Fisher–Rao std. dev.	0.0262	0.0270
H_1 bars	9	13
Max H_1 lifetime / cutoff	4.3%	3.9%

Remark: A Genuinely Unexpected Result

The delimited zone's aggregate entropy, 7.450 bits, differs from the whole-country figure, 7.494 bits, by only 0.044 bits – an order of magnitude smaller than the 0.734-bit range spanned by the twenty-one whole-country pools of Volume II, Chapter 9. This was not anticipated: the working assumption motivating this entire volume, stated in the preface above and repeated throughout Volume II's warnboxes, is that a whole-country pool's aggregate statistics are potentially quite different from those of any properly delimited substrate zone within it, because the whole-country pool additionally mixes in whatever other substrate layers the rest of the country contributes. For Turkey and this particular eleven-province interior delimitation, that additional mixing turns out to move the aggregate entropy by less than 0.05 bits – a genuinely informative negative result, not a confirmation of the motivating concern. Two explanations are available, and the data at hand cannot yet distinguish them: either (a) Turkey's toponymic substrate is comparatively homogeneous at the bigram-entropy level across its coastal and interior provinces, so that whole-country and delimited-zone estimates coincide more closely here than the concern anticipated in general; or (b)

this particular eleven-province delimitation, despite excluding the coastal fringes, is still broad enough (population 9,492 pool entries, 11 provinces) to retain most of the whole-country pool's internal heterogeneity. Distinguishing these requires at minimum a second, more narrowly delimited zone from the same country – exactly the kind of further case study Section 2.3 proposes as the immediate next step, rather than a conclusion this single case study is positioned to draw on its own.

The persistent-homology comparison in Table 2.2's last two rows reproduces, once more and at delimited-zone resolution, this series' now-familiar finding: both pools show short-lived H_1 bars at 3–5% of the filtration's own scale, squarely inside the noise-floor range Volume II, Chapter 5's stability theorems license reading as sampling artefact rather than genuine topological structure.

2.3 Roadmap: Further Zones Reachable from Data Already at Hand

The following delimitations are, in principle, immediately reachable using the self-verification protocol of Section 1.2 and the national dump files already available from Volume II, without requiring any further data acquisition. None of the numbers below has been computed; each is listed here as a concrete, falsifiable next step for a future chapter of this volume, precisely so that this roadmap cannot later be mistaken for a set of results.

1. **A second, narrower Anatolian zone — resolved, see Chapter 3.** Repeating Section 2.1's protocol with the four-province Cappadocia core (Nevşehir, Niğde, Kayseri, Aksaray) directly tests explanation (b) of the notebox above, and the result is reported in full in the next chapter.

2. **An Alpine German-speaking zone from `AT.txt` and `CH.txt` jointly — resolved, see Chapters 4 and 5.** Both the canton-level self-verification protocol and the elevation-field alternative sketched below were carried out in full: the former delimits a seventeen-canton German-speaking core and compares it against both the whole-country Switzerland pool and a combined Austria+German-Switzerland pool (Chapter 4); the latter splits the combined Austria+Switzerland pool at a 1,000 m elevation threshold using GeoNames' own `elevation` field, and compares the two delimitation criteria directly (Chapter 5).

3. **A Maghreb coastal-versus-interior split — resolved, see Chapter 9.** `DZ.txt`, `MA.txt`, and `TN.txt` were each split into coastal and interior pools following Section 1.2's protocol. The whole-country Maghreb clustering of Volume II, §9.3 survives the split: each country's own coastal and interior Fréchet means are closer to each other, in Fisher–Rao distance, than either is to any same-type pool in a different country. The entropy gap between coastal and interior itself, however, has no consistent sign across the three countries.

4. **Formal comparison against *Toponomastic Entropy*'s original zone boundaries**, wherever a documented definition for the relevant zone is available, so that this

volume's province-level delimitations can be checked directly against the parent series' own boundaries rather than only against each other.

5. **Feature-class decomposition of the entropy gap — resolved, see Section 3.4.** Chapter 3's closing notebox left open whether Cappadocia's higher aggregate entropy reflects genuine substrate signal or simply a higher proportion of feature-class T (hypsographic/natural-feature) entries relative to feature-class P (populated places) in a landscape as distinctive as Cappadocia's. The decomposition rules out the composition explanation: Cappadocia's class-T share is not elevated relative to the other two pools, and the entropy gap is present within both feature classes independently, more strongly so in class T.

One of the five original roadmap items remains open (the formal comparison against *Toponomastic Entropy*'s published zone boundaries); the other four are resolved within this volume, as noted against each item above. That remaining item is stated at the level of "here is exactly what data and protocol step would be needed, and here is exactly what question answering it would resolve," which is the standard this volume holds itself to throughout.

Chapter 3

Case Study Two: Narrowing to a Four-Province Cappadocia Core

3.1 Zone Definition and Motivation

Chapter 2's closing notebox left two explanations open for why the eleven-province Central Anatolia delimitation's aggregate entropy (7.450 bits) sat so close to the whole-country Turkey figure (7.494 bits): either Turkey's toponymic substrate is comparatively homogeneous across coastal and interior provinces, or the eleven-province pool was itself still too broad to reveal the heterogeneity a properly delimited zone would show. This chapter tests the second explanation directly, by repeating Section 1.2's protocol on a four-province subset of the same eleven: Nevşehir, Niğde, Kayseri, and Aksaray, the historical Cappadocia core, using the admin1 codes already verified in Section 2.1 (50, 73, 38, 75 respectively — no new verification step was needed, since these four codes were already recovered as a by-product of Chapter 2).

3.2 Pool Composition and Aggregate Statistics

Table 3.1 reports the pool contribution of each of the four provinces; Table 3.2 places the resulting statistics alongside both of Chapter 2's figures for direct three-way comparison.

Table 3.1: Cappadocia core: pool contribution by province. Total delimited pool: 1,544 entries.

Province	Entries
Kayseri	768
Aksaray	291
Nevşehir	253
Niğde	232

Table 3.2: Three-way comparison: whole-country Turkey, the eleven-province Central Anatolia delimitation, and the four-province Cappadocia core, all computed by the identical pipeline.

Quantity	Turkey (whole)	Central Anatolia (11)	Cappadocia (4)
Pool n	70,260	9,492	1,544
Mean name length	9.18	9.81	9.44
Aggregate H_2 (bits)	7.494	7.450	7.647
ΔH_2 vs. whole country	—	-0.044	$+0.153$
Fisher–Rao mean (n=40)	0.2370	0.2448	0.2310
Fisher–Rao min	0.147	0.160	0.000
H_1 bars	13	9	13
Max H_1 lifetime / cutoff	3.9%	4.3%	3.1%

Methodological Transparency: A Zero Fisher–Rao Distance, Explained

Table 3.2's Fisher–Rao minimum of 0.000 for Cappadocia is not a computational error: the $n = 40$ sample (seed 42, the fixed convention used throughout this series) drew the toponym *çamlıca* ("pine-grove place," a common Turkish place-name element) twice, from two different underlying GeoNames entries in two different villages, producing two bigram-identical, hence Fisher–Rao-identical, sample points. This is a real, expected consequence of sampling $n = 40$ points without replacement from a pool of only 1,544 entries in a language with a small number of highly productive toponymic formation patterns: common name elements recur across nearby villages far more often, proportionally, in a small pool than in a 70,260-entry whole-country pool. It is reported here rather than silently re-seeded away, because it is itself informative about the small-pool regime this chapter's narrower delimitation enters — exactly the kind of effect a whole-country analysis is too large to ever encounter.

3.3 Discussion: Explanation (b) Is Supported

The three-way comparison in Table 3.2 resolves Chapter 2's open question, at least for this one country: entropy divergence from the whole-country baseline *grows*, not shrinks, as the delimitation narrows. The eleven-province pool differed from the whole country by -0.044 bits; the four-province pool differs by $+0.153$ bits, more than three times as large a deviation in the opposite direction. This is exactly the signature explanation (b) predicted: the eleven-province Central Anatolia delimitation, despite excluding Turkey's coastal fringes, was still broad enough to retain most of the whole-country pool's statistical behaviour, and only the considerably narrower Cappadocia core reveals a toponymic substrate whose bigram entropy is genuinely, measurably higher than either the eleven-province or whole-country estimate.

Remark: What "Higher Entropy" Does and Does Not Mean Here

Consistent with this whole series' repeated caution on this exact point (Volume II, §9.4; Chapter 2's own warnbox above), Cappadocia's higher aggregate entropy is not, on its own, evidence of richer or more diverse toponymic *substrate layering* in a historical-linguistic sense. Two considerably more mundane candidate explanations have not been ruled out with the data at hand: first, Cappadocia's distinctive volcanic-tuff landscape (the "fairy chimney" formations the region is known for) plausibly supports an unusually large number of distinct micro-toponyms for small natural features — entries of feature class T specifically — relative to a typical Anatolian province, which would inflate bigram diversity through referential richness rather than substrate mixing; second, at $n = 1{,}544$, this pool is more than an order of magnitude smaller than the eleven-province one, and the aggregate-entropy estimator's own finite-sample bias (not separately quantified in this chapter) has not been ruled out as a contributor to the $+0.153$-bit gap. Distinguishing "genuine substrate signal" from "feature-class composition effect" from "finite-sample bias" would require, at minimum, recomputing Table 3.2's entropy column separately for feature class P and feature class T within each of the three pools — a natural next step, not carried out here, and recorded as an addition to the roadmap of Section 2.3.

The Fisher–Rao and persistent-homology rows of Table 3.2 show a less clear-cut pattern than the entropy row: Cappadocia's mean Fisher–Rao distance (0.231) sits between the whole-country and eleven-province figures rather than outside their range, and its H_1-lifetime ratio (3.1%) is if anything slightly *lower* than the other two, not higher — both consistent with this series' running finding that the $n = 40$, $K = 729$ Fisher–Rao and persistent-homology summaries are considerably less sensitive to zone delimitation than the aggregate corpus-level entropy is (a pattern already remarked on for the eighteen whole-country pools in Volume II, §9.2, and now confirmed once more across a delimitation boundary rather than only across countries). The clean three-way entropy result and the comparatively flat Fisher–Rao/TDA results should therefore both be reported together, rather than the entropy result alone: they jointly indicate that whatever this narrower delimitation is picking up, it shows up specifically in the aggregate bigram distribution's shape and not in the fine-grained geometry or topology of individual $n = 40$-toponym samples drawn from it.

3.4 Feature-Class Decomposition: Settling the Open Question

The notebox above left two candidate explanations for Cappadocia's $+0.153$-bit entropy gap undistinguished: a feature-class composition effect (an unusually high share of class-T natural-feature entries, inflating diversity through referential richness) versus a genuine substrate or finite-sample effect. Table 3.3 resolves this by recomputing aggregate H_2 separately for feature classes P (populated places) and T (hypsographic/natural features) within each of the three pools.

Table 3.3: Aggregate entropy decomposed by feature class, for all three Turkish pools of this chapter.

Pool	n_P	P share	n_T	$H_2(P)$	$H_2(T)$
Turkey (whole)	54,604	77.7%	15,656	7.507	7.030
Central Anatolia (11)	6,826	71.9%	2,666	7.525	6.991
Cappadocia (4)	1,207	78.2%	337	7.713	7.576

Remark: The Composition Explanation Is Not Supported

Table 3.3's second column rules out the feature-class-composition explanation directly: Cappadocia's class-T share, 21.8%, is not elevated relative to the other two pools — it sits between the eleven-province figure (28.1%) and the whole-country figure (22.3%), and is in fact *lower* than Central Anatolia's. The volcanic-tuff landscape's distinctive natural features do not translate into an outsized proportion of class-T entries in this delimitation. What Table 3.3 shows instead is that Cappadocia's entropy is higher in *both* feature classes independently: $H_2(P) = 7.713$ against 7.507 (whole country) and 7.525 (eleven-province), and, even more strikingly, $H_2(T) = 7.576$ against 7.030 and 6.991 — a 0.55-to-0.59-bit gap in the natural-feature class alone, several times larger than the 0.15-bit gap in the aggregate figure. Whatever is driving Cappadocia's elevated aggregate entropy, it is not a shift in the P/T mixture; it is present within both classes taken separately, and is markedly stronger in class T than in class P. This is consistent with – though does not on its own prove – the substrate-signal explanation over the composition explanation, and narrows the remaining open question to a choice between genuine substrate diversity and the finite-sample bias flagged in the notebox above, particularly for class T, whose Cappadocia-pool count ($n_T = 337$) is small enough that finite-sample effects remain a live concern the data at hand cannot rule out.

Chapter 4

The Alpine German-Speaking Core: Switzerland and Austria

4.1 A Verified Linguistic Delimitation

Switzerland's cantons are self-reported in `CH.txt` with short alphabetic admin1 codes (Section 1.2's verification protocol applied here recovers, for instance, `ZH` for Zürich, `GE` for Genève, `TI` for Ticino, each read directly from that canton's own `ADM1` row). Switzerland's four national languages are distributed unevenly across its twenty-six cantons: seventeen are treated here as unambiguously German-speaking, four are French-speaking (Vaud, Neuchâtel, Genève, Jura), one is Italian-speaking (Ticino), one is officially trilingual (Graubünden, German/Romansh/Italian), and two further cantons – Valais and Fribourg – are officially bilingual with a French-majority population, and Bern, though German-majority overall, contains the French-speaking Jura bernois district. This chapter delimits a "German-speaking core" using only the seventeen cantons with no official French, Italian, or Romansh-speaking population of any size (Table 4.1), explicitly excluding all nine of the above – Vaud, Neuchâtel, Genève, Jura, Valais, Fribourg, Ticino, Graubünden, and Bern – rather than assigning any of them by approximation.

Methodological Transparency: Why Bilingual Cantons Are Excluded, Not Approximated

Bern, Valais, Fribourg, and Graubünden each contain genuine German-speaking populations and toponymic layers alongside their French, Romansh, or Italian ones, at a finer (municipal or district) resolution than this chapter's admin1-level protocol can separate without a further, unverified assumption about which municipalities belong to which linguistic sub-region. Rather than approximate this split, all four cantons – together with the four unambiguously French-speaking cantons and Italian-speaking Ticino – are omitted entirely from the delimited pool. This makes the delimited pool a conservative, high-confidence German-speaking core, smaller than the true German-speaking population of Switzerland (which the Federal Statistical Office puts at somewhat over 60% of the national total), not an attempt to reconstruct that figure exactly.

4.2 Pool Composition and Comparison to the Whole-Country Baseline

Table 4.1 reports the pool contribution of each of the seventeen included cantons; Table 4.2 compares the resulting delimited pool against the whole-country Switzerland pool of Volume II, §9.1, and against a combined Austria+German-Switzerland pool (Austria's own whole-country pool, already essentially monolingual German at the national level, is added without further internal delimitation).

Table 4.1: German-speaking Swiss cantons: pool contribution. Total delimited pool: 10,561 entries across seventeen cantons.

Canton	Entries
Zürich (ZH)	3,766
St. Gallen (SG)	1,118
Uri (UR)	836
Thurgau (TG)	833
Luzern (LU)	808
Schwyz (SZ)	595
Glarus (GL)	475
Nidwalden (NW)	439
Aargau (AG)	434
Obwalden (OW)	293
Solothurn (SO)	284
Basel-Landschaft (BL)	181
Zug (ZG)	158
Schaffhausen (SH)	123
Appenzell Ausserrhoden (AR)	106
Appenzell Innerrhoden (AI)	81
Basel-Stadt (BS)	31

Table 4.2: Switzerland (whole country), the seventeen-canton German-speaking delimitation, and the combined Austria+German-Switzerland pool. Whole-country CH figures reproduced from Volume II, Table 9.4.

Quantity	CH (whole)	CH (German, 17)	AT+CH (German)
Pool n	23,959	10,561	41,650
Mean name length	10.44	10.46	10.24
Aggregate H_2 (bits)	7.882	7.653	7.667
ΔH_2 vs. CH whole	—	-0.229	-0.215
Fisher–Rao mean (n=40)	0.2280	0.2416	0.2548
H_1 bars	7	6	4
Max H_1 lifetime / cutoff	3.8%	2.2%	5.5%

Remark: Excluding Non-German Cantons Lowers Entropy Substantially

Restricting Switzerland's whole-country pool to its seventeen unambiguously German-speaking cantons lowers aggregate entropy by 0.229 bits ($7.882 \to 7.653$) – more than five times the magnitude of the Central Anatolia delimitation's shift in Chapter 2 (-0.044 bits), and comparable in magnitude to the Cappadocia shift ($+0.153$ bits) though opposite in sign. This is the pattern this whole volume has been built to look for: a whole-country pool that mixes several distinct linguistic strata (German, French, Italian, Romansh, in Switzerland's case) shows measurably higher aggregate entropy than a pool restricted to a single one of those strata, exactly as the mixing intuition motivating Volume II, Chapter 9's warnboxes predicted. Unlike the Turkish case studies, where a purely geographic delimitation (interior provinces) produced only a small shift, here a delimitation aligned with a genuine, independently known linguistic boundary (canton-level language majority) produces a shift several times larger – circumstantial but suggestive evidence that *linguistic* delimitation, where it can be verified, moves the aggregate-entropy needle more than geographic delimitation alone.

Combining the German-Swiss pool with Austria's whole-country pool (itself already essentially monolingual German at the aggregate level) produces a $41{,}650$-entry combined pool with $H_2 = 7.667$, close to but not identical to the German-Switzerland-only figure (7.653) – Austria's own whole-country entropy (7.626, Volume II, Table 9.1) is close enough to German-Switzerland's that combining the two shifts the aggregate only modestly. The Fisher–Rao and persistent-homology rows show the same broadly flat pattern already familiar from every other chapter in this series: mean Fisher–Rao distance and H_1-lifetime ratio move within the same narrow bands regardless of delimitation, while aggregate entropy is the quantity that responds most visibly to linguistic composition.

4.3 Cross-Reference to Chapter 5

A natural further question – whether the German-speaking Alpine core's entropy is better explained by *linguistic* delimitation (this chapter) or by *elevation* (a topographically Alpine, rather than linguistically German, criterion) – is addressed directly in the next chapter, which repeats a version of this comparison using the GeoNames digital-elevation field rather than admin1 codes.

Chapter 5

Elevation as a Data-Native Delimitation Criterion

5.1 Motivation: A Delimitation Requiring No Admin Lookup at All

Every delimitation so far in this volume has relied on the admin1 self-verification protocol of Section 1.2. GeoNames national dumps carry a further field not yet used in this series for delimitation purposes: column 16, `elevation` (a self-reported elevation in metres for each entry, distinct from column 17's coarser `dem` raster-derived estimate), populated independently of any administrative boundary. This chapter uses that field directly, splitting the combined Austria+Switzerland pool (both countries' entire national dumps, not restricted to any linguistic subset) into a high-elevation and a low-elevation half, to ask whether an elevation-based "Alpine" delimitation identifies a comparably coherent zone to Chapter 4's linguistically delimited one.

5.2 Coverage Caveat

Methodological Transparency: The `elevation` Field Is Not Populated for Every Entry

Of the combined $55{,}048$-entry Austria+Switzerland feature-class-P/T pool, only $13{,}879$ entries (25.2%) carry a valid, non-placeholder `elevation` value; the remainder have the field empty or set to the GeoNames placeholder -9999 and were excluded from this chapter's analysis entirely, rather than imputed. This coverage rate is stated here explicitly because it materially limits what this chapter's split can support: the $13{,}879$-entry analysable pool is a non-random subset of the full pool (entries with recorded elevation likely differ systematically from entries without one, most plausibly toward smaller or more remote settlements and named natural features where elevation was thought worth recording), and no correction for this selection effect has been attempted.

5.3 The Elevation Split

The 13,879-entry analysable pool was split at 1,000 m elevation: 12,134 entries (87.4%) at or above the threshold, 1,745 (12.6%) below it. Table 5.1 reports the resulting statistics, alongside the country-level breakdown of the split.

Table 5.1: Elevation split of the combined Austria+Switzerland analysable pool at 1,000 m.

Quantity	High ($\geq$ 1,000 m)	Low ($<$ 1,000 m)
Pool n	12,134	1,745
of which, Austria	3,221	402
of which, Switzerland	8,913	1,343
Mean name length	11.23	9.60
Aggregate H_2 (bits)	7.871	7.867
Fisher–Rao mean (n=40)	0.2551	0.2361
H_1 bars	4	6
Max H_1 lifetime / cutoff	2.6%	2.4%

Remark: Elevation Moves Name Length, Not Entropy

The two elevation bands differ by 0.004 bits in aggregate entropy — statistically and practically indistinguishable, an order of magnitude smaller than either Chapter 4's linguistic-delimitation shift (0.229 bits) or Chapter 3's narrowing shift (0.153 bits) – despite a substantial and clearly real difference in mean name length (11.23 versus 9.60 characters, a gap larger than several of the whole-country comparisons in Volume II, Table 9.1). High-elevation toponyms in this combined pool are genuinely, systematically longer than low-elevation ones – consistent with the general pattern, visible informally across many of this series' Alpine and mountain examples, that natural-feature names at altitude (ridges, passes, peaks) tend to carry more descriptive compound elements than settlement names in lowland valleys – but this does not translate into a detectably different bigram-entropy profile. This is, once again, this series' now-familiar caution applied to a new case: raw name length and aggregate bigram entropy are separate quantities (Volume II, §9.2's finding, re-confirmed here across an elevation boundary rather than a national one), and a delimitation criterion that moves one need not move the other.

5.4 Elevation Versus Linguistic Delimitation: A Direct Comparison

Table 5.2 places this chapter's elevation split alongside Chapter 4's linguistic delimitation, both computed from overlapping underlying country pairs (Austria and Switzerland), to ask directly which criterion identifies a more entropy-distinctive zone.

Table 5.2: Elevation-based versus linguistically delimited Alpine zones, both drawn from Austria and/or Switzerland.

Delimitation	Criterion type	Pool n	$\lvert\Delta H_2\rvert$ (bits)
CH German cantons (Ch. 4)	Linguistic (admin1)	10,561	0.229
Alps $\geq 1{,}000$ m (this chapter)	Elevation	12,134	0.008
Alps $< 1{,}000$ m (this chapter)	Elevation	1,745	0.012

The linguistic row's ΔH_2 is relative to Switzerland's whole-country aggregate ($H_2 = 7.882$, Volume II, Table 9.4). The two elevation rows' ΔH_2 values are relative to the full 13,879-entry analysable pool's own aggregate ($H_2 = 7.879$), *not* to Table 9.1's whole-country figures, since the elevation-split baseline is itself a coverage-restricted subset (Section 5.2) and not directly comparable to an unrestricted whole-country pool.

Methodological Transparency: An Honest Answer: Linguistic Delimitation Wins Here, but the Comparison Is Not Fully Fair

On the evidence in Table 5.2, linguistic delimitation identifies a far more entropy-distinctive zone than elevation does, for this particular country pair and this particular 1,000 m threshold. This should not, however, be read as a general finding that administrative/linguistic criteria are intrinsically more informative than physical-geographic ones: the two delimitations are not measuring comparable things (one separates language communities, the other separates terrain), the elevation comparison's baseline is itself a coverage-restricted subset (Section 5.2), and a single threshold value (1,000 m) was tested rather than a range. A fairer test would compare elevation-based delimitation against a *geographic*, not linguistic, whole-country baseline, or would vary the elevation threshold to check whether some other cutoff produces a larger split – neither of which this chapter attempts, and both of which are recorded as roadmap items below.

Chapter 6

The British Isles: Four Constituent Nations and a Celtic Fringe

6.1 A Delimitation Requiring No Interpretation at All

Unlike every delimitation so far in this volume, `GB.txt`'s admin1 field requires no numeric-code recovery step: its four values are the self-explanatory strings `ENG`, `SCT`, `WLS`, and `NIR`, each verified, as in every other chapter, against its own `ADM1`-feature-class row (England, Scotland, Wales, and Northern Ireland respectively). This chapter treats each of the four as its own delimited pool, adds Ireland's whole-country pool from Volume II, and constructs a fifth, combined "Celtic fringe" pool (Scotland+Wales+Northern Ireland+Ireland, deliberately excluding England) to ask whether the whole-country United Kingdom pool's entropy – already the single highest of the twenty-one whole-country pools in Volume II, Table 9.1 – is a genuine property of any one of its constituent parts, or an artefact of combining all four.

6.2 Six Pools, One Country Group

Table 6.1 reports aggregate statistics for the four constituent nations individually, for Ireland as a whole (already computed in Volume II but reproduced here for direct comparison), for the Celtic-fringe combination, and for the whole United Kingdom pool of Volume II, Table 9.1.

Remark: Every Part Is Lower Than the Whole

This is the cleanest instance in this volume of the pattern first identified in Chapter 4 for Switzerland: the whole-country United Kingdom pool's aggregate entropy, 7.950 bits, is higher than *every one* of its four constituent nations taken individually (range 7.717–7.992, and note that even Northern Ireland's 7.992, the single highest individual figure, is still below the whole-UK figure). It is higher even than the Celtic-fringe combination (7.894), which itself combines three of the four constituent nations plus Ireland and so already captures most of the cross-national mixing effect – adding England to that combination raises the aggregate a further 0.056 bits, to the whole-UK figure. Mixing more distinct national/regional toponymic strata together monotonically raises this series' aggregate-

Table 6.1: The British Isles, decomposed. "Whole UK" reproduces Volume II, Table 9.1; every other row is computed fresh in this chapter, either as a single admin1-delimited pool (Scotland, Wales, Northern Ireland, England) or as an explicit combination (Celtic fringe).

Pool	n	Mean len.	H_2	H_1 bars	Lif./scale
Scotland	21,755	11.33	7.717	8	3.4%
Wales	6,490	9.85	7.881	10	2.3%
Northern Ireland	1,005	9.93	7.992	5	4.4%
England	38,289	9.81	7.847	14	5.5%
Ireland (whole)	16,129	10.60	7.751	17	7.3%
Celtic fringe (Sc.+Wa.+NI+Ire.)	45,379	10.83	7.894	10	4.3%
Whole UK (Volume II, Table 9.1)	67,588	10.30	7.950	7	2.5%

entropy statistic, at every scale tested so far in this volume: four cantons versus seventeen (Chapter 4), four nations versus one (this chapter), and, more weakly, eleven provinces versus one (Chapter 2). This is not a surprising result in retrospect – a Shannon-entropy-of-a-mixture is well known, in general, to be at least as large as the pool-size-weighted average of the component entropies, plus a non-negative Jensen-Shannon-type mixing term, whenever the component distributions differ at all – but it is the first time in this series that the effect has been shown this cleanly, across five nested and overlapping real pools computed from a single self-consistent protocol.

Remark: A Formal Note on Why This Was Expected, in Hindsight

The pattern in Table 6.1 is, in fact, a direct consequence of a standard identity for Shannon entropy of a mixture distribution: if a pool is the union of m sub-pools with weights $w_1, \ldots, w_m$ (proportional to sub-pool size) and per-sub-pool bigram distributions $q_1, \ldots, q_m$, then the aggregate entropy of the mixture $\bar{q} = \sum_i w_i q_i$ satisfies $H(\bar{q}) \geq \sum_i w_i H(q_i)$, with equality only when every q_i is identical. The gap between the two sides is exactly the Jensen–Shannon-type mixing term this series has not previously named explicitly, and it is this term, not any change in the component distributions themselves, that Table 6.1's "every part is lower than the whole" pattern reflects. This does not make the pattern any less genuine an empirical finding about this particular set of real GeoNames pools – the mixing term's magnitude depends entirely on how different the four nations' real bigram distributions actually are, which is an empirical fact about the data, not a mathematical guarantee of any particular size – but it does mean the pattern should be read as a direct, expected consequence of pooling distinct distributions together, not as a newly discovered substantive claim about British toponymy specifically.

Fisher–Rao and persistent-homology statistics in Table 6.1 continue this volume's now-thoroughly-confirmed pattern of showing far less sensitivity to delimitation than aggregate entropy: every lifetime/scale ratio in the table, from Wales's 2.3% to Ireland's 7.3%, remains inside or close to the noise-floor range this series' stability theorems (Volume II, Chapter 5) license reading as sampling artefact, regardless of

which of the seven pools in Table 6.1 supplied the underlying $n = 40$ sample.

Chapter 7

Ireland's Four Historic Provinces

Methodological Transparency: A Field That Already Encodes the Provinces, Not the Counties

`IE.txt`'s `admin1` field does not encode Ireland's twenty-six counties; a direct scan of every `A/ADM1` row returns exactly four names – Connacht, Leinster, Munster, and Ulster, the four historic provinces – each with its own single-letter code. This is a coarser first-order division than the county-level delimitations this volume has built for other countries by hand-picking individual `ADM1` or `ADM2` rows; here, no picking is required or possible, since the file's own top administrative tier already is the province. Chapter 6's "Ireland (whole)" row ($n = 16{,}129$, Table 6.1) is the sum, within a small non-adversarial discrepancy documented below, of the four pools this chapter delimits individually.

7.1 Four Provinces, Self-Verified

Table 7.1 reports all four `ADM1` rows recovered from `IE.txt` by `find_admin1_code`, applied without exception to every province the file itself declares.

Table 7.1: Ireland's four historic provinces, self-verified as `ADM1` rows in `IE.txt`.

Province	Self-reported name	Code
Connacht	Connacht	C
Leinster	Leinster	L
Munster	Munster	M
Ulster	Ulster	U

Remark: Ulster Here Is a Rump, by Construction of the Data, Not by Any Choice of This Chapter

Historic Ulster comprises nine counties; six – Antrim, Armagh, Down, Fermanagh, Londonderry, and Tyrone – fall within Northern Ireland and so within `GB.txt`, already delimited as its own pool in Chapter 6. Only the remaining three – Cavan, Donegal, and Monaghan – lie within the Republic of Ireland and so within `IE.txt`'s `admin1=U` rows. The `admin1=U` pool this chapter delimits is there-

fore a genuine rump of the historic province, not the province as a whole; this is a fact about where the Irish border falls, recovered directly from which counties `IE.txt` does and does not contain, not an editorial choice about how to split Ulster.

7.2 Pool Composition and Aggregate Statistics

Table 7.2 reports n, mean normalised name length, aggregate bigram entropy H_2 (Laplace $\varepsilon = 1$, $K = 729$), and Fisher–Rao persistent-homology statistics ($n = 40$, seed$= 42$, Lidstone $\alpha = 0.5$) for each of the four provinces, computed by the unchanged pipeline of Appendix A and Volume II, Appendix 8.2–8.3, applied to `IE.txt` at sub-national resolution for the first time in this volume.

Table 7.2: Ireland's four historic provinces. Columns as in Table 6.1.

Province	n	Mean len.	H_2	H_1 bars	Lif./scale
Connacht	4,913	10.09	7.705	13	5.0%
Leinster	4,448	10.78	7.753	10	2.2%
Munster	4,873	11.12	7.768	5	4.1%
Ulster (rump)	1,882	10.15	7.793	10	2.4%
Sum of four	16,116	—	—	—	—
Ireland (whole, Table 6.1)	16,129	10.60	7.751	17	7.3%

Remark: A Thirteen-Row Discrepancy, of the Same Small, Non-Adversarial Kind Already Documented for Austria

The four provinces sum to 16,116, thirteen fewer than the 16,129 recorded for the whole-country pool in Table 6.1. `load_country_pool` restricts only to feature classes `P` and `T` and a normalised name of length at least three, exactly as `load_delimited_pool` restricted by `admin1` does; the discrepancy is not attributable to any difference in filtering criteria between the two functions. It matches, in both direction and order of magnitude, the three-row discrepancy Section **??**'s companion warning box documented for Austria and Chapter 16's own three-row discrepancy for Pakistan: a small number of rows carrying a blank or non-standard `admin1` field that `load_country_pool`'s whole-file scan still counts but no single-province `admin1`-restricted scan can, by construction, capture. No attempt is made here to track down these thirteen rows individually; the discrepancy is reported, not resolved, in keeping with this volume's standing commitment to honest small-count reporting over silent reconciliation.

Remark: Ulster's Entropy Is the Highest of the Four, But Only Marginally

Ulster's rump pool, at $n = 1{,}882$ the smallest of the four provinces by a wide margin (under half of Leinster's, the next smallest), also carries the highest aggregate entropy of the four, 7.793. This is the same direction of effect Chapter 19 reports for Valle d'Aosta ($n = 844$, the smallest pool in that chapter's table, also carrying

its highest entropy) and Chapter 18 reports for Assam – small pools in this series do, repeatedly, tend toward the upper end of their comparison group's entropy range. The effect here, however, is far weaker than in either of those two cases: Ulster's 7.793 exceeds Munster's 7.768, the next-highest of the three larger Irish provinces, by only 0.025 bits, whereas Valle d'Aosta's excess over Veneto is more than half a bit and Assam's elevation above its comparison group is of a broadly similar order. Ulster's small-pool entropy elevation is real but modest; it does not "stand out" from the other three provinces' range the way the two more extreme cases do, and pool size alone is accordingly not treated in this series as a sufficient explanation for any single elevated entropy figure, Ulster's included.

7.3 Roadmap

1. **A cross-border Ulster comparison.** This chapter delimits only the three-county rump of Ulster within the Republic; `GB.txt`'s six Northern Irish counties, already available as the single `Northern Ireland` pool in Table 6.1, could be combined with this chapter's `admin1=U` pool into a nine-county reunified-Ulster pool, directly testing whether the historic province coheres toponymically across the modern political border the way Chapter 15's cross-border Istrian comparison tested Slovenian against Croatian Istria.

2. **Fréchet-distance clustering of the four provinces.** No pairwise Fréchet-mean distance matrix has yet been computed among Connacht, Leinster, Munster, and Ulster; the same `pairwise_frechet_distances` routine Section 10.5 first executed in this volume could establish whether Ulster's small-pool profile translates into geometric isolation from the other three, or whether, as its modest entropy gap here suggests, it remains close to them in Fisher–Rao space.

7.4 Reproducibility: What Was Actually Run

Every figure in this chapter reuses `load_country_pool`, `load_delimited_pool`, `find_admin1_code`, `aggregate_entropy`, `sample_and_fr_matrix`, and `persistence` exactly as Appendix A and Volume II, Appendix 8.2–8.3 define them, applied to `IE.txt` for the first time at sub-national resolution in this volume. No `admin2`-level extraction was required or possible: all four provinces in Table 7.1 are the file's own top-level `ADM1` rows, recovered directly by `find_admin1_code`, and the whole-Ireland baseline row in Table 7.2 reproduces Table 6.1 exactly, confirming that no parameter drifted between the two chapters' independent runs of `load_country_pool` against the same file.

Chapter 8

Resolving the Ireland Roadmap: A Reunified Ulster Comparison

Methodological Transparency: Scope of This Chapter

Section 7.3's first item proposed combining Chapter 7's three-county Ulster rump (`IE.txt`, `admin1=U`, $n = 1{,}882$) with Chapter 6's six-county Northern Ireland pool (`GB.txt`, `admin1=NIR`, $n = 1{,}005$) into a single nine-county reunified-Ulster pool spanning the modern political border. That combination is carried out here. The second roadmap item – a full Fréchet-distance clustering of all four Irish provinces – remains open and is not attempted in this chapter.

8.1 A Nine-County Pool Spanning Two Country Files

`load_delimited_pool` was applied unchanged to each of `GB.txt` (restricted to `admin1=NIR`) and `IE.txt` (restricted to `admin1=U`), and the two resulting name lists were concatenated with no further filtering. This is the first pool in either volume of this series assembled from two country files at once; no new code was required beyond Python's list concatenation, since both source lists were already normalised, `P/T`-only, length-≥ 3 `asciiname` strings produced by the identical function.

Table 8.1: The reunified nine-county Ulster pool against its two constituent parts. Columns as in Table 6.1.

Pool	n	Mean len.	H_2	H_1 bars	Lif./scale
Northern Ireland (Table 6.1)	1,005	9.93	7.992	5	4.4%
Ulster rump (Table 7.2)	1,882	10.15	7.793	10	2.4%
Reunified Ulster (9 counties)	2,887	10.07	7.805	12	6.2%

Remark: The Whole Is Not Simply Between the Two Parts

Reunified Ulster's aggregate entropy, 7.805, sits close to but above the rump's 7.793 and well below Northern Ireland's 7.992 – broadly consistent with the mixing-entropy identity of Section 6.2's second notebox, since Northern Ireland

alone carries the higher of the two component entropies and the combined pool's aggregate necessarily reflects both components' distributions together. What that identity does not by itself predict is the persistence figure: the reunified pool's lifetime/scale ratio, 6.2%, is higher than *both* of its constituent parts (4.4% and 2.4%), the same qualitative pattern Chapter 6 flagged for Ireland-whole against its own four constituent nations, and a reminder that this series' stability theorems bound how much a persistence ratio can move under resampling noise, not that combining two pools must place the combination's ratio between its parts.

8.2 A Direct Cross-Border JSD

`bigram_distribution` and `jsd` (Section 14.6) were applied directly to the Northern Ireland and Ulster-rump pools, without first building the reunified pool, to ask how toponymically distinct the two sides of the border are from each other: $\mathrm{JSD}(\mathrm{NI}, \mathrm{rump}) = 0.0305$. This falls inside the middle band of Table 21.1's survey – higher than any pairing among Ireland's own four (all-Irish) provinces, comparable to Chapter 11's Talysh-versus-Azerbaijan figure (0.0447), and below the Samtskhe-Javakheti-versus-Georgia figure (0.0706). A century of separate administration and, for Northern Ireland, a different `asciiname`-generating GeoNames source file entirely, corresponds to a real but moderate toponymic divergence by this measure – larger than the divergence among the Republic's own provinces, but not of the same order as the Caucasus zones' more pronounced linguistic-minority signal.

8.3 Reproducibility: What Was Actually Run

Every figure in this chapter reuses `load_delimited_pool`, `aggregate_entropy`, `sample_and_fr_matrix`, `persistence`, `bigram_distribution`, and `jsd` exactly as Appendix A and Section 14.6 define them, applied here for the first time to a pool spanning two country files (`GB.txt` restricted to `admin1=NIR`, and `IE.txt` restricted to `admin1=U`) concatenated before the aggregate-statistics and Fisher–Rao steps. Both recomputed constituent pool sizes, $1{,}005$ and $1{,}882$, matched Chapters 6 and 7's reported figures exactly.

Chapter 9

A Maghreb Coastal-Versus-Interior Split

9.1 Extending the Protocol to Three Further Countries

Volume II, §9.3 found that Algeria, Morocco, Mauritania, and Tunisia cluster together coherently at whole-country resolution, among the seventeen-country extension's tightest geographic groupings. This chapter asks whether that coherence is a genuine Maghreb-wide toponymic signal or an artefact of comparing whole, internally mixed national pools, by applying Section 1.2's self-verification protocol to split each of Algeria, Morocco, and Tunisia individually into a coastal and an interior pool. (Mauritania is omitted from this chapter: its entire populated coastline is a single short stretch around Nouakchott, too narrow a case for the same kind of split.)

Methodological Transparency: The Coastal/Interior Label Is a Declared Geographic Judgement, Not a Field in the Data

Unlike this volume's admin1-code recovery, which is mechanically verified from each province's own `ADM1` row, no field in the GeoNames dump states whether a given province touches the sea. The coastal/interior assignment below was made from ordinary geographic knowledge of each country's coastline (which provinces border the Mediterranean or the Atlantic) exactly as Chapter 2 declared its eleven-province Central Anatolia boundary: openly, from outside the data, and not claimed to be recovered from it. What *is* recovered from the data, mechanically and exactly as elsewhere in this volume, is each named province's own self-reported admin1 code (Section 1.2, step 1) and the resulting pool composition once a coastal/interior province list has been supplied.

Algeria's population/toponym file self-reports 62 distinct admin1 codes; fourteen of them – Tlemcen, Aïn Témouchent, Oran, Mostaganem, Chlef, Tipaza, Alger, Boumerdès, Tizi Ouzou, Béjaïa, Jijel, Skikda, Annaba, and El Tarf – are the country's Mediterranean-facing wilayas and form the coastal set; the remaining forty-six named wilayas form the interior set, and two further codes (`00`, blank) covering 110 pool entries correspond to no self-reported province name and are excluded from both. Morocco's twelve post-2015 regions split five into a nine-region coastal set (Tanger-Tétouan-Al Hoceïma, Oriental, Rabat-Salé-Kénitra, Casablanca-Settat,

Marrakech-Safi, Souss-Massa, Guelmim-Oued Noun, Laâyoune-Sakia El Hamra, and Dakhla-Oued Ed-Dahab – the last two in Western Sahara) and a three-region interior set (Fès-Meknès, Béni Mellal-Khénifra, Drâa-Tafilalet); a further handful of legacy/duplicate admin1 codes covering 59 pool entries (0.07% of the whole-country pool) are excluded as unclassifiable. Tunisia's twenty-four governorates split ten coastal (Tunis, Ben Arous, Nabeul, Bizerte, Sousse, Monastir, Mahdia, Sfax, Gabès, and Médenine) against fourteen interior (Ariana, Manouba, Zaghouan, Béja, Jendouba, Le Kef, Siliana, Kairouan, Kasserine, Sidi Bouzid, Kébili, Tataouine, Gafsa, and Tozeur), with the single placeholder code `00` (130 entries) excluded.

9.2 Results

Table 9.1 reports the full pipeline's output for both halves of each of the three countries, alongside each country's whole-pool aggregate entropy from Volume II, Table 9.1, for reference.

Table 9.1: Coastal-versus-interior split for Algeria, Morocco, and Tunisia. "Whole" reproduces Volume II, Table 9.1.

Country	**Pool**	n	Mean len.	H_2	H_1 bars	Lif./scale
Algeria	Coastal	5,524	11.92	7.402	7	4.5%
Algeria	Interior	19,122	12.21	7.335	10	7.5%
Algeria	Whole	24,756	12.14	7.354	–	–
Morocco	Coastal	55,960	11.61	7.466	17	4.0%
Morocco	Interior	23,056	11.49	7.545	14	4.4%
Morocco	Whole	79,075	11.58	7.500	–	–
Tunisia	Coastal	2,434	12.06	7.523	9	2.0%
Tunisia	Interior	3,588	12.52	7.399	13	5.5%
Tunisia	Whole	6,152	12.34	7.426	–	–

Remark: No Consistent Sign Across the Three Countries

Algeria and Tunisia agree: in both, the coastal pool's aggregate entropy is higher than the interior pool's (7.402 against 7.335 for Algeria, a $+0.067$-bit gap; 7.523 against 7.399 for Tunisia, $+0.124$ bits), and higher than the whole-country figure in both cases. Morocco disagrees: its interior pool's entropy, 7.545, is *higher* than its coastal pool's, 7.466 – the opposite sign, by a comparable magnitude (-0.079 bits coastal-minus-interior). No single "coastal toponyms are more/less diverse than interior ones" rule holds across all three countries in this delimitation. This is itself the honest finding: whatever coastal/interior structure this particular criterion picks up, it is not a uniform Maghreb-wide effect, and a genuinely different candidate explanation would be needed for each country individually before any general claim could be made — exactly the caution this series has applied to every other delimitation criterion tested so far (Chapters 2 and 4, and Volume II, §9.4).

9.3 Does the Whole-Country Maghreb Clustering Survive the Split?

The question motivating this chapter – whether Volume II's whole-country Maghreb cluster reflects real cross-national coherence or is an artefact of pooling internally heterogeneous countries – is answered directly by comparing each country's coastal Fréchet mean against its own interior Fréchet mean and against the other two countries' coastal and interior means, six points in total. Table 9.2 reports, for each of the six pools, its single nearest neighbour by Fisher–Rao geodesic distance among the other five.

Table 9.2: Nearest neighbour, by Fréchet-mean Fisher–Rao distance, among the six coastal/interior pools of this chapter.

Pool	**Nearest neighbour**	Distance
Algeria, coastal	Algeria, interior	0.0401
Algeria, interior	Algeria, coastal	0.0401
Morocco, coastal	Morocco, interior	0.0406
Morocco, interior	Morocco, coastal	0.0406
Tunisia, coastal	Tunisia, interior	0.0426
Tunisia, interior	Tunisia, coastal	0.0426

Remark: National Identity Dominates the Coastal/Interior Split

Every one of the six pools' single nearest neighbour, out of the five other pools available, is its *own* country's other half, not any same-type (coastal-to-coastal or interior-to-interior) pool in a different country. Algeria's coastal and interior means are 0.040 apart; the next-closest pool to Algeria's coastal mean, of the remaining four, is 0.044 away (Tunisia's interior mean, from Table 9.1 row order not shown in full here) – close, but still second. This answers this chapter's motivating question directly: the whole-country Maghreb clustering of Volume II, §9.3 is not an artefact of coastal/interior heterogeneity being mistaken for cross-national coherence. National-level toponymic signal remains the dominant clustering axis even after this finer delimitation splits each country in two – a country's coastal and interior toponym pools resemble each other, in Fisher–Rao geometry, more than either resembles the same geographic stratum in a neighbouring country. Combined with §9.2's finding that the coastal/interior entropy *gap* itself has no consistent sign across the three countries, the overall picture is that country-level identity is the robust signal in this delimitation exercise, and coastal/interior structure, where it exists at all, is a smaller and country-specific effect superimposed on it.

This chapter resolves the fourth item of the Chapter 2 roadmap not listed as closed there: the Maghreb coastal/interior split has now been carried out for all three countries with a real coastline suited to it, using the same self-verification protocol as every other delimitation in this volume, and the whole-country clustering result it was designed to test has survived the finer split intact.

Chapter 10

The Caucasus: A Three-Country Reproducibility Check, an Elevation Split, and a Cross-National Clustering Test

10.1 Motivation: A Region With No Internal Admin1 Split to Verify

Georgia, Armenia, and Azerbaijan differ from every zone delimited so far in this volume in one structural respect: each is already its own national dump file, so there is no internal admin1 boundary within a single file comparable to Chapter 2's Turkish provinces or Chapter 4's Swiss cantons. A linguistic sub-national delimitation is not out of reach in principle – Georgia's Javakheti region and Azerbaijan's Talysh-speaking south-east are both real, documented minority-language areas – but assembling such a delimitation responsibly would require external linguistic sourcing this volume has not gathered, and is left as an open item in Section 10.6 rather than attempted here from guesswork. What *is* reachable immediately, from the three national dump files already used elsewhere in this series plus `TR.txt`, is threefold: a reproducibility check of the three countries' whole-pool statistics against Volume II's published figures; an extension of Chapter 5's data-native elevation criterion to a second, considerably taller mountain range; and a whole-country clustering test in the spirit of Volume II, §9.3's Maghreb finding, benchmarked against Turkey. This chapter goes beyond the five-item roadmap of Section 2.3, which did not anticipate it; it is undertaken because `TR.txt` was already at hand for Chapter 2 and the three Caucasus files were already available from Volume II, §9.1, making the comparison free of any new data acquisition.

10.2 Reproducibility Check: Recomputing Three Whole-Country Pools

Table 10.1 recomputes Georgia's, Armenia's, and Azerbaijan's whole-country pools from scratch, using exactly the extraction and normalisation code of Appendix A, extended to these three files (reproduced in condensed form in Section 10.7 below), and compares the result against the figures already published in Volume II (Georgia and Azerbaijan: Table 9.1; Armenia: §7.1's original three-zone table).

Table 10.1: Independent recomputation of three whole-country pools against Volume II's published figures. "This chapter" uses the same national dump files and the same pipeline; agreement confirms the pipeline is stable and the underlying data files have not changed between volumes.

Country	Pool n	H_2 (this chapter)	H_2 (Vol. II)	Δ
Georgia	6,737	7.463	7.463	0.000
Armenia	6,823	7.417	7.418	−0.001
Azerbaijan	6,923	7.733	7.733	0.000

Remark: An Exact, Not Approximate, Reproduction

All three pool sizes match Volume II's exactly, and two of the three entropy values match to the third decimal place; Armenia's differs by 0.001 bits, well within floating-point rounding of a 729-term sum and not indicative of any change to the underlying file or pipeline. This is a stronger reproducibility result than any within-volume check this series has run so far, precisely because it is a genuine independent recomputation from the raw file rather than a citation of an earlier table – the kind of check this volume's warnboxes have repeatedly said matters more than any single reported number.

10.3 Coverage Caveat: Elevation Reporting Is Far Sparser Here Than in the Alps

Methodological Transparency: A Second, Much Lower, Coverage Rate

Of the combined 20,483-entry Georgia+Armenia+Azerbaijan pool, only 1,259 entries (6.15%) carry a valid, non-placeholder `elevation` value – roughly a quarter of Chapter 5's 25.2% rate for Austria+Switzerland, and the shortfall is not evenly spread: Georgia's own coverage ($757/6{,}737 = 11.2\%$) is nearly three and four times Armenia's ($218/6{,}823 = 3.2\%$) and Azerbaijan's ($284/6{,}923 = 4.1\%$) respectively. GeoNames' `elevation` field is evidently populated at a rate that varies by an order of magnitude across national dumps for reasons this series has no data to explain (likely differences in each country's own contributing survey data rather than anything about the terrain itself); this is recorded here as a limitation on what any elevation-based delimitation drawn from these three files in particular

can support, not generalised into a claim about GeoNames' elevation field as a whole.

Where the analysable Alpine pool clustered near a natural 1,000 m threshold, the 1,259-entry Caucasus elevation-populated pool is centred far higher: minimum 0 m, maximum 5,201 m, mean 2,059.7 m, median 2,167 m – unsurprising given the Greater Caucasus range, but a reminder that the fixed 1,000 m threshold Chapter 5 used for the Alps is not a neutral, region-independent cutoff: applied here it would place 78.5% of the analysable pool in the "high" band. Both a 1,000 m split (for direct comparability with Chapter 5) and a median split at 2,167 m (to keep the two bands close to equal size) are reported below rather than choosing one and discarding the other.

10.4 The Elevation Split, Two Ways

Table 10.2: Elevation split of the combined Georgia+Armenia+Azerbaijan analysable pool ($n = 1{,}259$), at the region-derived median and at the fixed 1,000 m threshold of Chapter 5.

Band	Pool n	Mean length	H_2 (bits)	H_1 bars (lif. %)
$\geq$ 2,167 m (median)	630	11.77	7.651	15 (4.9%)
$<$ 2,167 m (median)	629	11.69	7.549	11 (3.8%)
$\geq$ 1,000 m	988	11.72	7.518	6 (1.7%)
$<$ 1,000 m	271	11.75	7.799	9 (3.7%)

Methodological Transparency: The Sign of the Elevation Effect Depends on Where the Threshold Is Placed

At the median split, the high band's entropy exceeds the low band's by 0.102 bits (7.651 against 7.549) – the same direction as Chapter 5's Alpine result, though roughly twenty-five times larger in magnitude (0.102 against 0.004 bits). At the fixed 1,000 m threshold, the sign *reverses*: the low band's entropy exceeds the high band's by 0.281 bits (7.799 against 7.518), the largest single elevation-related gap reported anywhere in this volume, in the opposite direction from the median split computed on the very same underlying pool. This is not a contradiction in the data; it is a direct demonstration of why Chapter 5's closing caution – that a single threshold value was tested rather than a range – was the right one to state. The 271-entry low-band pool at the 1,000 m split is small and composed of whichever populated places and natural features happen to sit below that particular line in this particular region; a different threshold draws a materially different pool from the same 1,259 raw entries, and this chapter reports both splits rather than the one that tells a cleaner story.

Mean name length, by contrast, is essentially threshold-invariant here (11.69–11.77 characters across all four bands) and well above the combined pool's

unrestricted-pool average – consistent with Chapter 5's finding that elevation-recorded entries skew toward longer, more descriptive natural-feature names regardless of exactly where the elevation cutoff falls. All four H_1 lifetime-versus-scale ratios (1.7%–4.9%) remain inside the noise-floor range this series has used throughout (Volume II, Chapter 5): none of the four bands shows persistent topological structure distinguishable from sampling noise.

10.5 Cross-National Clustering: Do the Three Countries Cohere as a Group?

Volume II, §9.3 found Algeria, Morocco, Mauritania, and Tunisia clustering together at whole-country resolution, and Chapter 9 of this volume found that the same clustering, at *sub*-national resolution, is dominated by national identity rather than surviving as a coastal/interior signal. This section asks the whole-country version of that question for a different region: do Georgia, Armenia, and Azerbaijan's whole-country Fréchet means sit closer to each other, in Fisher–Rao distance, than any of the three sits to a fourth, geographically adjacent but linguistically distinct comparator – Turkey, whose whole-country pool is already available from Chapter 2.

Table 10.3: Pairwise Fréchet-mean Fisher–Rao distances (radians, $n = 40$ sample, seed 42) among the four whole-country pools.

	Georgia	**Armenia**	**Azerbaijan**	**Turkey**
Georgia	–	0.0553	0.0551	0.0623
Armenia	0.0553	–	0.0522	0.0644
Azerbaijan	0.0551	0.0522	–	0.0590
Turkey	0.0623	0.0644	0.0590	–

Remark: A Coherent Trio, and an Unforced Comment on Turkey's Own Nearest Neighbour

Each of Georgia's, Armenia's, and Azerbaijan's nearest neighbour among the other three pools is one of the other two Caucasus countries, not Turkey: Georgia's nearest is Azerbaijan (0.0551, against 0.0553 to Armenia); Armenia's and Azerbaijan's nearest is each other (0.0522). Every within-trio distance (0.0522–0.0553) is smaller than every Caucasus-to-Turkey distance (0.0590–0.0644), a clean separation with no overlap. This parallels, at whole-country resolution, the Maghreb clustering Volume II reported for four different countries, and – unlike Chapter 9's finding for the Maghreb at sub-national resolution – this chapter does not test whether the signal survives a finer split, since, as Section 10.1 states, no verified sub-national delimitation is available yet for these three countries. One further point is offered only as an observation, not a finding this data can support on its own: Turkey's own nearest neighbour among the three Caucasus pools is Azerbaijan (0.0590), the two countries whose national languages belong to the same Turkic family – a pattern consistent with what is independently known about their linguistic relationship, but a single pairwise distance from an

unrestricted, mixed-substrate whole-country pool is far too weak a basis to treat this as a confirmation of anything, and it is reported here in the same spirit as Chapter 2's open questions: a concrete, falsifiable observation for a future, more careful chapter to test properly.

Remark: A Data Quirk Worth Recording: Armenia's Repeated Toponym

Armenia's $n = 40$ Fisher–Rao sample (the same seed-42 draw underlying Table 10.3) happened to select the same normalised name, `jradzor`, twice from two distinct GeoNames entries, producing a pairwise Fisher–Rao distance of essentially zero (2.5×10^{-7}, floating-point noise around a true value of 0) and pulling that pool's minimum pairwise distance to 0.0000 in a way none of this chapter's other three whole-country samples exhibit. This is not a pipeline error: `AM.txt` genuinely contains more than one feature named "Jradzor" at different coordinates, a real and unremarkable repetition of a common Armenian place-name element rather than a duplicate row. It is recorded here because a reader cross-checking Armenia's pairwise-distance summary statistics against the other three countries' would otherwise have no way to know why its minimum is qualitatively different (exactly 0 rather than merely small).

10.6 Roadmap Addendum

Two concrete next steps follow directly from this chapter and are recorded here rather than pre-empted with invented results, in the same spirit as Section 2.3:

1. **A verified sub-national delimitation within one Caucasus country.** Georgia's Samtskhe-Javakheti region and Azerbaijan's Talysh-speaking Lankaran/Astara district are both documented minority-language areas that could, in principle, be delimited by the admin1 self-verification protocol of Section 1.2 once their constituent province names are identified from a reliable source external to this series – a step this chapter has deliberately not taken from memory alone.

2. **A finer elevation-threshold sweep.** Rather than the two threshold choices reported in Section 10.4, a sweep across the full 0–$5{,}201$ m range, tracking how the entropy gap's sign and magnitude change continuously, would characterise the reversal noted above properly rather than documenting it at only two points.

10.7 Reproducibility: What Was Actually Run

Every number in this chapter was produced by the unchanged extraction, normalisation, entropy, Fisher–Rao, and Fréchet-mean routines of Appendix A and Volume II, Appendix 8.2, applied to `GE.txt`, `AM.txt`, `AZ.txt`, and `TR.txt` with two additions specific to this chapter: an elevation-column reader (column 16, matching Section 5.1's field choice, treating both an empty string and the placeholder -9999 as missing) and a median-threshold split alongside the fixed $1{,}000$ m one. No routine here is new in substance; both additions are thin wrappers around code already reproduced in full elsewhere in this volume.

Chapter 11

Resolving the Caucasus Roadmap: Two Verified Sub-National Zones and an Elevation-Threshold Sweep

Methodological Transparency: Scope of This Chapter

This chapter closes both open items of Section 10.6: a verified sub-national delimitation for two of the three Caucasus countries (Section 11.1), and a full elevation-threshold sweep replacing the two-point comparison of Section 10.4 (Section 11.3). No delimitation below is recovered from the data itself beyond the admin1-code self-verification already used throughout this volume; where a zone boundary reflects external ethnolinguistic knowledge rather than a field in the GeoNames file, this is stated openly, exactly as Chapter 9's coastal/interior split stated it.

11.1 Two Declared Sub-National Delimitations

Georgia: Samtskhe-Javakheti. Unlike every other country in this volume, Georgia's own first-order administrative division already includes, as a single self-reported `ADM1` unit, a region matching a documented linguistic minority area directly: Samtskhe-Javakheti (admin1 code `72` in `GE.txt`, recovered by the self-verification protocol of Section 1.2 with no external code table needed), a border region with a substantial and long-documented Armenian-speaking population dating from nineteenth-century resettlement. No judgement call was required to assemble this zone's province list, because it is already exactly one administrative unit; the only external fact brought in from outside the data is the region's documented linguistic-minority status, offered as motivation for why this particular one of Georgia's eleven `ADM1` units was chosen for this chapter rather than, say, Adjara or Guria.

Azerbaijan: a Talysh core. No single `ADM1` unit in `AZ.txt` corresponds to the Talysh-speaking area of the south-east; as in Chapter 9's coastal/interior split, a multi-rayon zone was assembled from ordinary geographic/ethnographic knowledge of where Talysh is spoken (the historic Talysh-Mughan lowlands and adjoining foothills), not from any field in the data: Lankaran Rayon (`29`), Lankaran Sahari (`30`, the city), Astara District (`08`), Lerik Rayon (`31`), Masally District (`32`), and Yardymli

Rayon (`66`). Each of the six codes was recovered from its own `ADM1` row exactly as Section 1.2 requires; the six-rayon *list* is a declared external judgement, openly stated, and offered as a reasonable first approximation rather than a claim of dialectological precision.

Methodological Transparency: Armenia Has No Equivalent Zone in This Chapter

Armenia is omitted from this chapter's sub-national delimitation because it lacks an equally documented, geographically compact linguistic-minority region comparable in scale to Samtskhe-Javakheti or the Talysh lowlands; assembling one would require a source this volume has not consulted, and inventing a plausible-sounding delimitation without one would violate this series' standing constraint against delimitations not traceable to either the data itself or an openly stated external source. Armenia's sub-national delimitation is therefore left as a genuinely open item, not resolved here.

11.2 Pool Composition and Results

Table 11.1: The two delimited zones against their respective whole-country baselines (Table 10.1).

Country	Pool	n	Mean len.	H_2	H_1 bars	Lif. %
Georgia	Samtskhe-Javakheti (delim.)	299	8.01	8.071	14	3.1%
Georgia	Whole country	6,737	9.61	7.463	6	2.8%
Azerbaijan	Talysh core (delim.)	710	8.72	7.949	14	7.3%
Azerbaijan	Whole country	6,923	9.31	7.733	22	2.3%

Remark: A Striking, Consistent Pattern – and Two Reasons Not to Over-Read It

Both delimited minority zones show *higher* aggregate entropy than their national baseline: Samtskhe-Javakheti by 0.608 bits (8.071 against 7.463), Talysh by 0.216 bits (7.949 against 7.733). Both gaps are far larger than Chapter 2's near-null 0.044-bit Central Anatolia result, and both run in the same direction – a genuinely suggestive pattern across the only two minority-region delimitations this series has computed. Two cautions apply before reading this as a general finding. First, $n = 2$ zones is not enough to establish a pattern; a single additional counter-example would overturn it, and none has been tested. Second, and more concretely, Samtskhe-Javakheti's pool ($n = 299$) is smaller than any other delimited zone in this volume (the next-smallest, Chapter 3's narrowest single province, exceeds it), and a 299-entry pool estimating a distribution over $K = 729$ bigram bins is a substantially thinner basis than this series' other aggregate entropy figures, even with Laplace smoothing. The direction of the effect is real in the data at hand; its magnitude and its generalisability to a larger sample from the same region are both open.

Remark: Name Length and Entropy Diverge Again, in a New Way

Both delimited zones have *shorter* mean name length than their national baseline (Samtskhe-Javakheti 8.01 against Georgia's 9.61; Talysh 8.72 against Azerbaijan's 9.31) while showing *higher* entropy – the opposite pairing from Chapter 5's Alpine finding, where the high-elevation band had *longer* names and statistically indistinguishable entropy. Between the two chapters, all four combinations of "longer/shorter" and "higher/lower/same entropy" have now been observed somewhere in this volume. This is, once again, this series' running caution rather than a new one: raw name length and aggregate bigram entropy are measuring different things, and neither predicts the other's direction from one delimitation to the next.

Talysh's H_1 lifetime-to-scale ratio, 7.3%, sits above every other ratio reported in this volume's real-data chapters (range elsewhere, Chapters 2–10: 1.7%–4.9%), which have been read throughout as inside the sampling noise floor licensed by Volume II, Chapter 5's stability theorems. Whether 7.3% from a single $n = 40$ draw marks a genuine departure from that floor or is itself still consistent with it is not settled by this chapter – it is recorded as an open question, alongside a concrete next step (a resampling check in the style of Volume II, §6.4, redrawing the Talysh $n = 40$ sample repeatedly and checking whether an H_1 bar this long recurs) rather than resolved either way here.

11.3 The Elevation-Threshold Sweep

Section 10.4 reported two points (a $1{,}000$ m and a median $2{,}167$ m split) with opposite-signed gaps and flagged a full sweep as the honest next step. Table 11.2 reports the aggregate entropy gap at eight thresholds spanning the full elevation range, together with a bootstrap standard error on each side's entropy estimate ($B = 500$ resamples with replacement of each band's own pool, seed $1/2$) and the resulting z-score for the gap.

Table 11.2: Elevation-threshold sweep, combined Georgia+Armenia+Azerbaijan analysable pool ($n = 1{,}259$). H_2 gap is high minus low; bootstrap standard errors from $B = 500$ within-band resamples; $z = \text{gap}/\sqrt{\text{se}_{\text{high}}^2 + \text{se}_{\text{low}}^2}$.

Thresh. (m)	n_{high}	n_{low}	$H_{2,\text{high}}$	$H_{2,\text{low}}$	Gap	se (h/l)	z
500	1,155	104	7.473	8.355	-0.882	0.019/0.039	-20.2
1,000	988	271	7.518	7.799	-0.282	0.021/0.033	-7.1
1,500	850	409	7.580	7.621	-0.041	0.022/0.030	-1.1
2,000	696	563	7.645	7.541	$+0.103$	0.023/0.027	$+2.9$
2,500	463	796	7.694	7.543	$+0.151$	0.027/0.024	$+4.2$
3,000	255	1,004	7.911	7.506	$+0.405$	0.031/0.022	$+10.8$
3,500	100	1,159	8.414	7.475	$+0.939$	0.041/0.020	$+20.6$
4,000	30	1,229	9.079	7.462	$+1.616$	0.031/0.020	$+43.8$

Remark: The Sign Flip Is Real and Sits at a Specific, Narrow Threshold Range

The gap crosses zero between $1{,}500\,\text{m}$ (gap -0.041, $|z| = 1.1$, indistinguishable from noise) and $2{,}000\,\text{m}$ (gap $+0.103$, $z = +2.9$, unlikely to be noise at this bootstrap resolution): the two splits reported in Section 10.4 ($1{,}000\,\text{m}$ and the $2{,}167\,\text{m}$ median) happened to land on opposite sides of this narrow crossing band, which is why they showed opposite signs. Below roughly $1{,}500$–$2{,}000\,\text{m}$, lower-elevation entries are more entropic; above it, higher-elevation entries are, and the effect grows sharply with each higher threshold tested, reaching $+1.616$ bits at $4{,}000\,\text{m}$. This is now a fully characterised, rather than merely observed, threshold dependence.

Methodological Transparency: The Largest Gaps Come From the Smallest, Most Extreme Pools, and the Bootstrap Does Not Fix This

The z-scores at the highest thresholds are large in part because the bootstrap standard error is computed *within* each fixed band – it quantifies how much H_2 would vary on repeated resampling *of the same* 30 *or* 100 *entries*, not how much it would vary on a fresh, independently drawn set of high-altitude Caucasus toponyms of the same size. A small, idiosyncratic pool (at $4{,}000\,\text{m}$, 30 entries, plausibly dominated by a handful of named peaks and passes with unusual, possibly non-native-substrate names) can have a large gap and a small within-sample bootstrap error simultaneously, without that gap being a stable property of "Caucasus toponyms above $4{,}000\,\text{m}$" in any generalisable sense. The z-scores in Table 11.2 should be read as evidence the observed gap is not an artefact of *this particular resampling procedure*, not as evidence it would replicate under a different or larger sample of the same population – a distinction Volume II, §6.4's resampling-experiment code was designed to test properly (redrawing from the underlying zone repeatedly, not bootstrapping a single fixed draw) and which this chapter's sweep, built for speed across eight thresholds, has not attempted.

11.4 What Remains Open

Both items of Section 10.6 are now resolved: a verified sub-national delimitation has been carried out for two of the three countries (Section 11.1), and the elevation-threshold sweep (Section 11.3) has replaced the earlier two-point comparison with a full characterisation of the sign flip. Three items follow directly and are left open rather than pre-empted:

1. An Armenian sub-national delimitation, once a suitable documented minority region and its constituent `ADM1` or `ADM2` units are identified from a source this volume has not yet consulted (Section 11.1's warnbox).

2. A proper resampling check of Talysh's elevated H_1 lifetime-to-scale ratio, following Volume II, §6.4's protocol rather than a single $n = 40$ draw.

3. An independent-redraw version of the elevation-threshold sweep at the two most extreme thresholds ($3{,}500\,\text{m}$ and $4{,}000\,\text{m}$), to test whether the large gaps reported

there are a stable population-level property or an artefact of the particular small pools GeoNames happens to report elevation for at those altitudes.

Chapter 12

Resolving the Roadmap: A Real-Data Resampling Check, an Independent-Redraw Elevation Test, and a First Verified Step Toward an Armenian Zone

Methodological Transparency: Scope of This Chapter

Section 11.4 left three items open. This chapter closes the first two in full and advances, without closing, the third: a resampling check of Talysh's H_1 lifetime-to-scale ratio (Section 12.1), an independent-redraw variance estimate for the elevation-threshold sweep's two most extreme thresholds (Section 12.2), and a first, name-verified but not yet pool-complete step toward an Armenian subnational delimitation (Section 12.3). All three sections use only the national dump files already in hand (`AZ.txt`, `GE.txt`, `AM.txt`) and the unchanged extraction, entropy, Fisher–Rao, and persistence routines of Appendix A and Volume II, Appendices 8.2–8.4; no new data acquisition beyond one external, openly-cited source for Section 12.3 was required.

12.1 A Resampling Check of Talysh's H_1 Lifetime-to-Scale Ratio

Section 11.2 flagged Talysh's 7.3% ratio, from a single $n{=}40$, seed-42 draw, as the highest reported anywhere in this volume's real-data chapters, and left open whether it marks a genuine departure from the sampling noise floor or is itself still consistent with it. Volume II, §6.4's protocol answers exactly this kind of question, but was built for a fitted *synthetic* Dirichlet model (Table **??**: $B{=}12$ independent redraws of the same generative distribution). Talysh's 710-entry pool is real GeoNames data, not a fitted model, so the redraws below draw $B{=}12$ independent $n{=}40$ subsamples *without replacement* directly from the verified 710-entry pool (seeds $0, 1000, 2000, \ldots, 11000$, matching `resampling_experiment`'s seed spacing), recomputing the Fisher–Rao distance matrix and H_1 persistence diagram for each draw

exactly as Section 11.2 did for the single seed-42 draw.

Table 12.1: Talysh, $B = 12$ independent n=40 draws without replacement from the verified 710-entry pool. Cutoff: $D_{\max}$ for that draw. Ratio: max H_1 lifetime / cutoff.

b	Seed	Cutoff	H_1 bars	Max lifetime (ratio)
0	0	0.3264	18	0.0337 (10.3%)
1	1000	0.2903	9	0.0161 (5.5%)
2	2000	0.3223	22	0.0106 (3.3%)
3	3000	0.3304	6	0.0091 (2.7%)
4	4000	0.3111	3	0.0083 (2.7%)
5	5000	0.3079	8	0.0199 (6.5%)
6	6000	0.3043	9	0.0220 (7.2%)
7	7000	0.3276	9	0.0159 (4.9%)
8	8000	0.2995	15	0.0160 (5.3%)
9	9000	0.3234	12	0.0114 (3.5%)
10	10000	0.3254	9	0.0078 (2.4%)
11	11000	0.2908	9	0.0079 (2.7%)
Mean across 12 draws		—	10.75	0.0149 (4.8%)
Original (seed 42, published)		0.3003	14	0.0219 (7.3%)

The 11 successive-draw bottleneck distances (Definition 6.2, computed via the same augmenting-path bipartite matching Appendix **??** specifies, here run via the open-source `persim` implementation rather than a bespoke one) range from 0.00385 to 0.01441, with mean $d_{\mathrm{B}}^{\mathrm{succ}} = 0.00743$, giving a 2×-mean noise threshold of 0.01487 – the same construction Table **??** used for the three synthetic zones.

Methodological Transparency: An Honest Result That Does Not Match Volume II's Pattern

All three of Volume II, §6.4's synthetic zones had their original canonical max lifetime fall *below* the 2×-mean noise threshold. Talysh's does not: the published seed-42 draw's max lifetime, 0.0219, exceeds the resampling-derived threshold, 0.0149, by a factor of 1.47. Ranked among the 12 resampling draws plus the original (13 values total), the original sits eleventh – higher than ten of the twelve independent redraws, and exceeded only by draw b=0 (0.0337, 10.3%) and draw b=6 (0.0220, barely). This does not settle the question Section 11.2 left open in Talysh's favour: a single ratio exceeding a noise threshold built from only 12 redraws is weak evidence on its own, and the redraws themselves are far from tightly clustered (ratios range 2.4%–10.3%, a spread of factor 4.3, itself larger than any single-country dispersion reported elsewhere in this volume). What the check does establish is that Talysh's original 7.3% figure is not an outlier artefact of one unlucky seed choice in the way a figure sitting comfortably inside the noise band would have been: it sits in the upper range of what repeated independent sampling from the same 710-entry pool actually produces, which is the honest, quantified version of "elevated" that Section 11.2 could not yet state precisely. Whether this reflects genuine topological structure in Talysh's toponymic substrate or simply

the wide resampling variance inherent to drawing $n{=}40$ points from a 710-entry, $K{=}729$-bin bigram space remains open, and is not resolved by this test alone.

12.2 An Independent-Redraw Variance Estimate for the Elevation Sweep's Two Most Extreme Thresholds

Section 11.3's warnbox noted precisely that its bootstrap standard errors quantify variability under resampling *the same fixed band with replacement*, not variability across a fresh draw of the underlying population – and that no fresh draw is available, since GeoNames' elevation field for the Caucasus is what it is. What *is* available, and was not yet tried, is subsampling *without* replacement: repeatedly drawing a 70% subset of each band's own entries ($B{=}500$ replicates, disjoint seeds per band), recomputing aggregate entropy on each subset. Unlike the bootstrap, no entry is ever duplicated within a draw, so the resulting spread is not inflated by the artificial repeated-point effect a with-replacement draw can introduce in a small pool; it remains, exactly as before, a property of resampling the *same fixed* 30- or 100-entry band, not evidence about a hypothetical larger population – a limitation this section states as plainly as Section 11.3 did, not one it claims to have solved.

Table 12.2: Independent-redraw (subsampling, 70%, without replacement, $B = 500$) variance estimate at the sweep's two most extreme thresholds, against Table 11.2's bootstrap figures.

Thresh. (m)	n_h, n_l	Gap	Bootstrap z	Subsamp. se (h/l)	Subsamp. z
3,500	100, 1,159	+0.939	+20.6	0.0240/0.0132	+34.3
4,000	30, 1,229	+1.616	+43.8	0.0142/0.0126	+85.0

Remark: The Gap Survives a Non-Bootstrap Test, With an Important Caveat Attached

Under subsampling, both gaps remain large relative to their standard errors – if anything more so than under the bootstrap ($z{=}34.3$ against 20.6 at 3,500 m; $z{=}85.0$ against 43.8 at 4,000 m), because subsampling without replacement tends to produce a tighter, less duplicate-inflated spread than bootstrapping the same small band with replacement. This rules out one specific concern: the gap is not an artefact of the bootstrap's own duplicate-resampling mechanic. It does *not* rule out the concern Section 11.3 actually raised, which was never about the bootstrap mechanic specifically but about whether *any* resampling of a fixed 30-entry pool can speak to population-level stability. Both the bootstrap and the subsampling estimate here draw exclusively from the same 30 (or 100) GeoNames rows; neither can distinguish "a stable property of high-Caucasus toponyms" from "a property of the specific 30 peaks and passes GeoNames happens to report elevation for above 4,000 m," because there is no larger set of high-Caucasus GeoNames entries to draw a genuinely fresh sample from. That distinction remains open, exactly as it was before this section, now with a second, non-bootstrap method agreeing with the first rather than resolving it.

Methodological Transparency: A Concrete, Testable Candidate Mechanism, Not Yet a Finding

One pattern in the raw data is worth recording as a hypothesis for a future chapter to test properly, not as an explanation this section argues for. Of the 30-entry 4,000 m-and-above pool, 22 entries (73.3%) contain the substring `mount`, `peak`, or `gora` (Russian for "mountain") in the original, pre-normalisation name – against 46.0% for the full 1,259-entry elevation-analysable pool, and a roughly monotonic rise across intermediate thresholds (41.4% at 2,000 m, 51.0% at 3,000 m, 62.0% at 3,500 m). The highest peaks in this region are disproportionately catalogued in GeoNames under English exonyms ("Mount Dzhangitau") or Russian survey-era names ("Gora Shkhara") rather than under a native-substrate name alone – unsurprising given these are the most prominent, most externally surveyed and mapped features in the dataset. Mixing English- and Russian-language generic elements into an aggregate bigram pool dominated, at the highest thresholds, by a mere 30 entries is a plausible mechanism for inflating that pool's bigram diversity (and hence its aggregate entropy) relative to lower bands where locally-named populated places dominate – but this is a candidate compositional explanation identified from the data at hand, not a demonstrated causal account, and no attempt is made here to separate this effect from any genuine substrate signal in the underlying native-language toponyms.

12.3 Toward an Armenian Delimitation: A Verified Yazidi Settlement List, Not Yet a Complete Zone

Section 11.1's warnbox stated that an Armenian sub-national delimitation would require "a source this volume has not consulted." This section consults one: English-language reference material on Armenia's Yazidi minority – the largest ethnic minority in Armenia, concentrated historically around Mount Aragats – names, at settlement resolution, twenty villages in Aragatsotn Province reported as almost entirely Yazidi-populated (Alagyaz, Arevut, Avshen, Charchakis, Ddmasar, Hako, Jamshlu, Kanch, Kaniashir, Metsadzor, Mijnatun, Mirak, Otevan, Rya Taza, Sadunts, Shamiram, Shenkani, Sipan, Sorik, Tlik) plus two Yazidi-majority villages in neighbouring Armavir Province (Yeraskhahun, Ferik), citing Omarkhali (2017) as the underlying source for the Aragatsotn list. This is a settlement-level list, not an administrative-code list, so it does not slot directly into the admin1 self-verification protocol of Section 1.2 the way Samtskhe-Javakheti's or Talysh's delimitations did; what this section does instead is verify each named settlement against `AM.txt` directly, entry by entry.

Table 12.3: Verification of the 22 named Yazidi-majority settlements against `AM.txt`. All 22 were found; none was assumed.

Province (admin1)	Villages named	Villages verified	Notes
Aragatsotn (`01`)	20	20	one homonym (Alagyaz, `admin1=05`) e
Armavir (`03`)	2	2	Yeraskhahun, Ferik
Total	22	22	

Methodological Transparency: What Was Verified, What Was Not, and Why This Is Not Yet a Zone

Every one of the 22 named settlements was located in `AM.txt` by exact normalised-name match against a `P`-class row with a non-zero recorded population, admin1 code `01` or `03` as appropriate; none was assumed or approximated. One genuine ambiguity was resolved by exclusion rather than guesswork: `AM.txt` contains a second, unrelated `Alagyaz` entry under admin1 code `05` (Kotayk), evidently a distinct feature from the Aragatsotn village of the same name, and was not included. This verification is, however, deliberately *not* carried through to a full delimited-zone pool in the style of Chapters 2–11, for a concrete reason: those zones were built by admin1 (or, for Talysh, multi-rayon admin1) restriction, which pulls in every `P/T` entry – populated places *and* natural features – within a verified administrative boundary, typically yielding pools in the hundreds. A list of 22 named villages is not an administrative boundary, and pooling only the 22 settlement names themselves gives a pool an order of magnitude thinner than Samtskhe-Javakheti's already-flagged n=299 (8.01 mean length off a pool with no natural features at all) – far too small to support this volume's standard pipeline, which itself draws n=40-point Fisher–Rao samples *from* a pool the size of this entire village list. No aggregate entropy, Fisher–Rao, or persistence figure is reported for this n=22 list for that reason, and none should be inferred from its absence.

Two paths would complete this delimitation properly, and are recorded here as the concrete next step rather than attempted under this chapter's time and data budget. First, `AM.txt` carries 82 `ADM2`-level community (*hamaynk*) rows for Armenia following the 2019 municipal reform; if the specific `ADM2` communities containing these 22 villages can be identified and verified by the same `find_admin1_code`-style lookup extended to `ADM2`, a community-level pool – still small, but plausibly in the low hundreds rather than 22 – could be built honestly. Second, and more conservatively, a wider read of Omarkhali (2017) or an equivalent primary source might extend the verified settlement list itself, since the sources consulted here name Aragatsotn and Armavir as the primary concentration but also mention smaller Yazidi populations in Ararat and Shirak provinces not itemised by settlement in the material reviewed for this section. Either route is left open rather than pre-empted with an approximated community boundary, in keeping with this volume's standing constraint.

12.4 Updated Roadmap

1. **The Armenian ADM2 lookup.** Identify and verify, by `ADM2`-code self-verification, which of `AM.txt`'s 82 community-level rows contain the 22 villages verified in Section 12.3, then build and report the resulting pool's statistics on the same footing as Chapters 2–11.

2. **A larger-B Talysh resampling check.** Section 12.1's B=12 draws matched Volume II, §6.4's synthetic protocol for direct comparability, but 12 real-data draws leave the 2.4%–10.3% ratio range poorly resolved; B=50 or more would tighten the noise-threshold estimate materially.

3. **Separating the generic-toponym effect from substrate signal.** Section 12.2's

`mount/peak/gora` compositional hypothesis is directly testable: recomputing the elevation-threshold sweep with such entries excluded would show whether the entropy gap survives, shrinks, or disappears once this candidate mechanism is controlled for.

12.5 Reproducibility: What Was Actually Run

Section 12.1 used `build_zone_codes/load_delimited_pool` (Appendix A) unchanged to recover the verified 710-entry Talysh pool, then `sample_and_fr_matrix` (Volume II, Appendix 8.2, item 4) with seeds $\{0, 1000, \ldots, 11000\}$ in place of the single seed 42, `ripser` for each draw's H_1 diagram, and the open-source `persim` package's bottleneck-distance routine (equivalent to Appendix **??**, item 7's augmenting-path matching, used here rather than reimplemented) for the 11 successive-pair distances. Section 12.2 reused Section 11.3's combined 1,259-entry elevation pool and `aggregate_entropy` routine unchanged, replacing the bootstrap's with-replacement resample by `numpy`'s `choice(..., replace=False)` at a fixed 70% subsample fraction, $B{=}500$ replicates per band, disjoint seed streams for the high and low bands. Section 12.3's verification is a direct linear scan of `AM.txt` for exact normalised-name matches against the 22-village list, with no statistical pipeline invoked. No routine in this chapter is new in substance beyond the subsampling replacement for the bootstrap and the settlement-name verification scan; both are thin, disclosed extensions of code already reproduced in full elsewhere in this volume.

Chapter 13

Resolving the Updated Roadmap: An Armenian ADM2 Zone and a Fifty-Draw Talysh Resample

Methodological Transparency: Scope of This Chapter

Section 12.4's updated roadmap left three items open. This chapter closes the first two: the ADM2-level Armenian delimitation (Section 13.1), completing the settlement-level verification of Section 12.3 into a full, statistics-bearing pool for the first time; and a B=50 extension of Section 12.1's B=12 Talysh resampling check (Section 13.2). The third item – separating the generic-toponym effect from substrate signal in the elevation sweep – remains open and is not attempted here. Both sections below use only `AM.txt` and `AZ.txt`, already in hand, and the unchanged pipeline of Appendix A and Volume II, Appendices 8.2–8.4.

13.1 Completing the Armenian Delimitation: From Twenty-Two Villages to a Six-Community ADM2 Pool

Section 12.3 verified all 22 named Yazidi-majority settlements against `AM.txt` but stopped short of building a delimited pool, since a 22-entry settlement list is an order of magnitude thinner than this volume's other pools and is not an administrative boundary. The first path that section left open – identifying the ADM2 (*hamaynk*, post-2019 community) rows that contain these 22 villages – is completed here. `find_admin1_code` (Appendix A) is extended, with no change to its refusal-over-guessing logic, to read column index 11 (`admin2 code`) in place of column index 10 (`admin1 code`); every one of the 22 villages' own rows in `AM.txt` already carries this code, so no new lookup source was required.

Remark: Alagyaz Resolves the Homonym Flagged in Section 12.3

The Aragatsotn village Alagyaz gives its name to Alagyaz Municipality (`13156511`), the community that also contains nine of the other nineteen Aragat-

Table 13.1: The six ADM2 communities containing the 22 verified Yazidi villages, self-reported in AM.txt. Every village's own row supplied its community code directly; none was inferred.

Province	Community (self-reported name)	Code	Villages it contains
Aragatsotn	Alagyaz Municipality	13156511	Avshen, Charchakis, Jamshlu, Kaniashir, Mijnatun, Mirak, Rya Taza, Sadunts, Shenkani, Sipan
Aragatsotn	Arevut Municipality	13156515	Arevut, Ddmasar, Hako, Kanch, Sorik, Tlik
Aragatsotn	Metsadzor Municipality	13156516	Metsadzor, Otevan
Aragatsotn	Shamiram Municipality	13156517	Shamiram
Armavir	Ferik Municipality	13156552	Ferik
Armavir	Metsamor Municipality	13156557	Yeraskhahun

sotn villages on the list – confirming, at the community level, that this is the same cluster Section 12.3 intended, and unrelated to the distinct admin1=05 (Kotayk) Alagyaz entry that section excluded on admin1 grounds alone.

Restricting AM.txt to these six community codes – every P/T entry within them, not only the 22 named villages themselves – yields a pool of **52** entries: the 22 named settlements plus 30 further populated places and named hydrological/hypsographic features sharing the same six communities, recovered mechanically by load_delimited_pool with no additional judgement call.

Table 13.2: The delimited six-community Armenian pool against the whole-country baseline. Columns as in Table 14.2.

Pool	n	Mean len.	H_2	FR min/mean/max	H_1 bars	Lif. %
Yazidi ADM2 zone (6 comm.)	52	7.58	9.031	0.119/0.212/0.266	14	5.6%
Armenia, whole country	6,823	—	7.417	—	—	—

Methodological Transparency: The Whole-Country Row Reuses Table ??, Not a Fresh Draw

Armenia's whole-country entropy (7.417) and pool size (6,823) are Volume II, Table **??**'s already-published figures for AM.txt, reused rather than recomputed, since load_country_pool applied to an unchanged file is deterministic; the FR/H_1 columns are left blank here because that table's n=40 sample used a different random draw than this chapter's zone-specific seed-42 draw and the two are not directly comparable cell-by-cell.

Remark: The Largest Entropy Gap in This Volume, From Its Smallest Full-Pipeline Pool

The six-community pool's aggregate entropy, 9.031 bits, exceeds Armenia's whole-country baseline by +**1.613** bits – larger than every other minority-zone gap this volume has computed (Samtskhe-Javakheti +0.608, Talysh +0.216, Sec-

tion 15.2's Slovenian Istria $+1.001$), and the largest of any delimitation in this series to date. It is also, at $n{=}52$, the smallest pool this volume has run its full entropy/Fréchet/persistence pipeline on – thinner even than Samtskhe-Javakheti's already-flagged $n{=}299$, and the same order of magnitude as the 22-village list Section 12.3 declined to run the pipeline on directly. The two observations sit in tension: this is simultaneously the volume's most striking gap and its least statistically supported one. A 52-entry pool estimating a $K{=}729$-bin bigram distribution, even Laplace-smoothed, is a substantially thinner basis than any other figure in Table 13.2's row, and the direction of the gap should be read as suggestive rather than established. The $n{=}40$ Fréchet-mean distance to the Armenia whole-country baseline, 0.0399 radians, sits within the range of other whole-country/zone distances reported elsewhere in this volume (Table 14.3: 0.0397–0.0566), which is itself notable: the entropy gap is the largest in the volume, but the Fréchet-distance gap is unremarkable by comparison – a dissociation between the two measures in the same spirit as Section 14.4's Fréchet/JSD disagreements, though here it is aggregate entropy rather than JSD diverging from the Fréchet picture.

13.2 A Fifty-Draw Extension of the Talysh Resampling Check

Section 12.1's $B{=}12$ check found the published seed-42 draw's H_1 lifetime-to-scale ratio, 7.3%, exceeded ten of twelve independent redraws, but flagged 12 draws as too few to resolve the 2.4%–10.3% noise range with any precision (Section 12.4, item 2). This section repeats the identical protocol – independent $n{=}40$ subsamples without replacement from the same verified 710-entry Talysh pool, seeds $0, 1000, 2000, \ldots, 49000$ – at $B{=}50$.

The 49 successive-draw bottleneck distances range from 0.00367 to 0.01441, mean $d_{\mathrm{B}}^{\mathrm{succ}} = 0.00735$, giving a $2\times$-mean noise threshold of 0.01469 – lower than the $B{=}12$ threshold of 0.01487 (Section 12.1), since a larger, more representative sample of successive-draw distances pulls the mean down only slightly here.

Remark: A Tighter Result in the Same Direction

The published seed-42 draw's max lifetime, 0.0219, still exceeds the $B{=}50$ noise threshold, 0.01469, by a factor of 1.49 – essentially unchanged from the $B{=}12$ check's factor of 1.47. What has changed is the ranking's precision: among all 51 values (the 50 redraws plus the original), the published draw now ranks **4th highest** – ahead of 47 of the 50 independent redraws – compared to $B{=}12$'s ranking of 11th of 13. A larger resampling budget has not overturned Section 12.1's finding that Talysh's original ratio is not an unlucky-seed artefact; it has sharpened it, moving the published draw from "above most redraws" to "among the highest few observed." This still does not by itself distinguish genuine topological substrate signal from wide inherent resampling variance ($B{=}50$'s ratio range, 0.6%–10.3%, remains a factor of 17 wide) – but a signal surviving a more than fourfold increase in resampling budget, in the same rank position relative to the

Table 13.3: Talysh, $B = 50$ independent n=40 draws without replacement from the verified 710-entry pool. Summary statistics across all 50 draws, plus the five highest and five lowest ratios individually. Columns as in Table 12.1.

b	Seed	Cutoff	H_1 bars	Max lifetime (ratio)
Five highest ratios				
0	0	0.3264	18	0.0337 (10.3%)
6	6000	0.3043	9	0.0220 (7.2%)
25	25000	0.2913	19	0.0210 (7.2%)
48	48000	0.3166	10	0.0220 (7.0%)
24	24000	0.2942	14	0.0201 (6.8%)
Five lowest ratios				
15	15000	0.3497	3	0.0022 (0.6%)
21	21000	0.3486	5	0.0068 (1.9%)
10	10000	0.3254	9	0.0077 (2.4%)
47	47000	0.3337	8	0.0082 (2.4%)
32	32000	0.3498	13	0.0086 (2.4%)
Mean across 50 draws		—	10.98	0.0133 (4.3%)
Original (seed 42, published)		0.3003	14	0.0219 (7.3%)

pool, is stronger evidence than the B=12 check alone provided.

13.3 Reproducibility: What Was Actually Run

Section 13.1 used the two-line `admin2` extension below, applied to `AM.txt`, in place of `find_admin1_code/load_delimited_pool`; every other routine (`aggregate_entropy`, `sample_and_fr_matrix`, `frechet_mean_sphere`, `persistence`) is Appendix A and Volume II, Appendix 8.2–8.3, unchanged. Section 13.2 reused Section 12.1's unchanged draw loop, extended from B=12 to B=50 seeds, and the same `persim` bottleneck routine for the 49 successive-pair distances.

```
# Extension used in Section 11.1: admin2-level analogues of
# find_admin1_code / load_delimited_pool (Appendix A), column index
# 11 (admin2 code) in place of column index 10 (admin1 code).
# Refusal-over-guessing behaviour is unchanged.
def find_admin2_code(path, settlement_name):
    with open(path, encoding='utf-8') as f:
        for line in f:
            parts = line.rstrip('\n').split('\t')
            if len(parts) < 12:
                continue
            if parts[6] == 'P' and normalize(parts[2]) == \
               normalize(settlement_name):
                return parts[11]          # admin2 code, opaque
                   string
    return None

```

```
def load_delimited_pool_admin2(path, admin2_codes):
    codes = set(admin2_codes)
    names = []
    with open(path, encoding='utf-8') as f:
        for line in f:
            parts = line.rstrip('\n').split('\t')
            if len(parts) < 12:
                continue
            if parts[6] not in ('P', 'T'):
                continue
            if parts[11] not in codes:
                continue
            az = normalize(parts[2])
            if len(az) >= 3:
                names.append(az)
    return names
```

Remark: Updated Roadmap

Of Section 12.4's three items, two are now closed by this chapter. The third – separating the `mount/peak/gora` generic-toponym effect from substrate signal in the elevation-threshold sweep – remains open and is recorded again here rather than attempted without a verified generic-toponym list, in keeping with this volume's standing constraint against approximated inputs to a quantitative claim.

Chapter 14

The Iberian Peninsula: Three Co-Official-Language Autonomous Communities

Methodological Transparency: Why This Chapter Uses Two New National Dumps

Every delimitation in this volume so far has drawn on the Caucasus, Anatolia, Alpine, British Isles, or Maghreb national dumps. This chapter opens two files not previously used anywhere in this series, `ES.txt` (Spain) and `PT.txt` (Portugal), and applies the unchanged admin1 self-verification protocol of Section 1.2 to a region chosen for a specific methodological reason stated in Section 14.1 below: unlike every sub-national zone this volume has delimited to date, the three regions studied here require *no* external judgement call about which provinces to group together.

14.1 A Delimitation With No Judgement Call

Talysh (Section 11.1) and the Maghreb coastal/interior split (Chapter 9) both required this volume to assemble a multi-province zone from ordinary geographic or ethnographic knowledge external to the data – an openly declared judgement call each time, but a judgement call nonetheless. Spain's own first-order administrative division removes that step entirely for three of its seventeen `ADM1` autonomous communities, each of which is already, by Spanish law, a single administrative unit with a co-official language alongside Castilian Spanish: the Basque Country (`Euskal Autonomia Erkidegoa`, recovered as its own `ADM1` row under that Basque-language name, not a Spanish or English gloss), Galicia, and Catalonia (`Catalunya`). Each region's admin1 code was recovered individually by Section 1.2's protocol, with no assumption that the codes follow any numeric pattern:

Remark: Why "No Judgement Call" Still Deserves a Caveat

Calling this delimitation judgement-free applies to *which provinces belong to the zone*, not to *which zones were chosen for this chapter*. Spain has seventeen autonomous communities; three were selected here because each hosts a documented co-official language distinct from Castilian Spanish (Basque, Galician,

Table 14.1: The three co-official-language autonomous communities, self-verified from their own `ADM1` rows in `ES.txt`.

Region	**Self-reported name in `ES.txt`**	**Code**
Basque Country	Euskal Autonomia Erkidegoa	59
Galicia	Galicia	58
Catalonia	Catalunya	56

Catalan), a fact brought in from outside `ES.txt` exactly as Section 11.1 brought in Samtskhe-Javakheti's documented Armenian-speaking population. The distinction that matters methodologically is narrower and more concrete: once a region is chosen, its `admin1` boundary already *is* the linguistically motivated zone, with no further external province list to assemble or defend – unlike Talysh, where the six-rayon list itself was an open judgement call even after each rayon's own name was externally supplied.

14.2 Pool Composition and Whole-Country Baselines

Table 14.2: The three delimited zones against the Spain and Portugal whole-country baselines. H_2: aggregate bigram entropy, $K = 729$, Laplace $\epsilon = 1$. FR: Fisher–Rao pairwise distance summary from the standard $n = 40$, seed 42 sample. H_1 bars and lifetime ratio from the same draw's Vietoris–Rips persistence diagram.

Pool	n	Mean len.	H_2	FR min/mean/max	H_1 bars	Lif. %
Basque Country	2,011	8.67	7.517	0.152/0.214/0.286	10	5.4%
Galicia	6,264	8.17	7.313	0.137/0.216/0.299	8	5.5%
Catalonia	3,655	11.73	7.351	0.142/0.240/0.329	13	3.8%
Spain, whole country	39,058	10.63	7.359	0.152/0.243/0.308	6	5.1%
Portugal, whole country	19,746	10.04	7.216	0.151/0.226/0.313	14	2.7%

Remark: Catalonia's Mean Length Is an Outlier in a Way the Other Two Regions Are Not

Catalonia's mean normalised name length, 11.73, is more than three letters longer than either Basque (8.67) or Galician (8.17), and also longer than the Spain-wide baseline (10.63) – the opposite pattern from Basque and Galicia, both of which are *shorter* than the national baseline. All three regions' aggregate entropies, by contrast, sit within a narrow 0.204-bit band (7.313–7.517), close to and statistically unremarkable against the Spain-wide figure (7.359). Name length and aggregate bigram entropy dissociating in this way is, once again, this series' standing caution rather than a new finding (Section 11.2 observed the same kind of dissociation, in the opposite direction, for Samtskhe-Javakheti and Talysh) – but this is the first chapter in the volume where the dissociation appears *within* a single country's three delimited zones rather than only between a zone and its baseline.

14.3 Cross-Regional Distances: A Genuine Galician–Portuguese Signal

Table 14.3: Pairwise Fréchet-mean Fisher–Rao distances (radians, $n = 40$ sample, seed 42) among all five pools.

	Basque	**Galicia**	**Catalonia**	**Spain**	**Portugal**
Basque	–	0.0461	0.0534	0.0566	0.0497
Galicia	0.0461	–	0.0506	0.0451	0.0397
Catalonia	0.0534	0.0506	–	0.0474	0.0556
Spain	0.0566	0.0451	0.0474	–	0.0487
Portugal	0.0497	0.0397	0.0556	0.0487	–

Remark: Galicia and Portugal Are Each Other's Nearest Neighbour

Galicia's nearest neighbour among the other four pools is Portugal (0.0397), not Spain as a whole (0.0451) and not either of the other two Spanish regions; symmetrically, Portugal's nearest neighbour among all four other pools is Galicia (0.0397), closer than its own country-to-country comparator, Spain (0.0487). This is a clean, mutual nearest-neighbour pair, and it lines up with an independently documented linguistic fact this series has not needed to invoke anywhere else: Galician and Portuguese descend from a shared medieval Galician-Portuguese language and remain, among Iberian Romance varieties, unusually close to one another. As with the Georgia–Armenia–Azerbaijan trio's clustering (Section 10.5) and Turkey–Azerbaijan's single nearest-neighbour link there, this is offered as a concrete, falsifiable observation consistent with known linguistic history, not as a claim that toponymic bigram statistics alone establish that history: a single pairwise Fréchet distance from unrestricted, mixed-substrate regional pools is weak evidence on its own, exactly as Section 10.5 cautioned for its own comparable finding.

Remark: Basque Is the Least Like Every Other Pool in the Table, Including Spain's Own Baseline

Basque's largest distance to any other pool $(0.0566$, to Spain whole-country$)$ is also the single largest entry anywhere in Table 14.3, and every one of Basque's four distances exceeds 0.046, higher than any distance among the remaining four pools except Catalonia–Portugal (0.0556). This is the pattern one would expect on independent linguistic grounds: Basque is a language isolate, with no demonstrated genetic relationship to Spanish, Galician, Catalan, or Portuguese, all four of which are Romance languages descended from Latin. A bigram-level toponymic signal picking up a language-family boundary this cleanly is a striking result, but the same caution applies as above – the underlying pools mix Basque-substrate and Spanish-substrate toponyms without separation (no attempt is made in this chapter to isolate toponyms of specifically Basque etymology from the wider Basque Country pool), so the finding should be read as consistent with,

not proof of, the isolate status that is independently well established.

14.4 A Population-Level Cross-Check: Jensen–Shannon Divergence

The Fréchet-mean distances above are built from a single $n = 40$ sample per pool and are, by this volume's own repeated caution (Section 12.1), subject to non-trivial sampling variance. As a population-level cross-check requiring no sampling at all, Table 14.4 reports the Jensen–Shannon divergence (base-2, in bits) between each pair of pools' *full* Laplace-smoothed bigram distributions – the same $K = 729$-bin construction `aggregate_entropy` uses internally, compared directly rather than reduced to a single entropy scalar.

Table 14.4: Pairwise Jensen–Shannon divergence (bits) between full, Laplace-smoothed ($\epsilon = 1$) bigram distributions – a population-level measure using every pool entry, not a $n = 40$ sample.

	Basque	**Galicia**	**Catalonia**	**Spain**	**Portugal**
Basque	–	0.1342	0.1796	0.1117	0.1887
Galicia	0.1342	–	0.1167	0.0444	0.0584
Catalonia	0.1796	0.1167	–	0.0567	0.1475
Spain	0.1117	0.0444	0.0567	–	0.0906
Portugal	0.1887	0.0584	0.1475	0.0906	–

Methodological Transparency: A Mechanical Confound the Fréchet-Distance Table Does Not Share

Under JSD, Galicia's nearest neighbour is Spain (0.0444), not Portugal (0.0584) – reversing Table 14.3's finding. The reason is not a disagreement about the underlying toponyms: it is a structural feature of population-level JSD that the sample-based Fréchet distance does not share. Galicia's 6,264-entry pool is a strict *subset* of Spain's 39,058-entry whole-country pool (every Galician toponym is also a Spanish one, by construction), so Spain's aggregate bigram distribution is mechanically pulled toward Galicia's simply by containing it, shrinking their JSD for reasons that have nothing to do with linguistic similarity. Portugal's pool shares no such containment relationship with Galicia's. Basque and Catalonia are subsets of Spain in exactly the same sense, and Table 14.4 shows the identical effect for them: Spain is each region's closest or second-closest neighbour by JSD, a relationship absent from the Fréchet-distance table (Section 14.3), which draws an equal-size $n = 40$ sample from each pool independently and is therefore not mechanically biased by whichever pool happens to contain another. This is a genuine methodological lesson rather than a data error: whole-population divergence measures and equal-size resampled distance measures answer subtly different questions when one pool nests inside another, and only the Galician–Portuguese link (Section 14.3) – the one pair with *no* subset relationship among

the two candidate explanations – survives under both methods unchanged.

14.5 Roadmap

1. **An admin2-level split of Catalonia.** Catalonia's outlying mean name length (Section 14.2) is reported here only in aggregate; `ES.txt`'s `admin2`-level *comarca* rows could support a finer-grained internal split (coastal versus Pyrenean comarques, for instance) in the style of Chapter 3's Cappadocia narrowing.

2. **A resampling check on the Basque–everything-else gap.** Section 14.3's Basque isolate finding rests on a single $n = 40$ draw per pool; Volume II, §6.4's protocol, already applied to Talysh in Section 13.2, could quantify how stable Basque's largest-distance ranking is under repeated independent sampling.

3. **Extending the JSD cross-check to every delimited zone in this volume.** Section 14.4's subset-containment caveat applies equally to every whole-country/subnational pair this volume has computed (Talysh against Azerbaijan, Samtskhe-Javakheti against Georgia, and so on); none of those chapters computed a population-level JSD cross-check alongside their Fréchet-distance figures, and this section's finding suggests they should.

14.6 Reproducibility: What Was Actually Run

Every figure in this chapter was produced by the unchanged `load_country_pool`, `load_delimited_pool`, `aggregate_entropy`, `sample_and_fr_matrix`, and `persistence` routines of Appendix A and Volume II, Appendix 8.2–8.3, applied to `ES.txt` and `PT.txt` for the first time in this volume, plus one addition: a `bigram_distribution` function returning the full Laplace-smoothed $K = 729$-bin probability vector *before* the final entropy reduction (`aggregate_entropy` calls this function internally already; Section 14.4 simply exposes its intermediate output and compares two such vectors via the standard Jensen–Shannon divergence, $\mathrm{JSD}(p, q) = \frac{1}{2}D_{\mathrm{KL}}(p\|m) + \frac{1}{2}D_{\mathrm{KL}}(q\|m)$, $m = \frac{1}{2}(p + q)$). No new zone-verification logic was required: all three regions in Table 14.1 are single `ADM1` units, recovered by `find_admin1_code` exactly as Section 4.2's Alpine cantons were.

```
# Addition used in Section 12.5: full bigram distribution (the
# intermediate quantity aggregate_entropy already computes, exposed
# directly for pairwise JSD comparison instead of reduced to
   entropy).
def bigram_distribution(names, K=729, eps=1.0):
    total = np.full(K, eps)
    for nm in names:
        total += bigram_vector(nm)
    return total / total.sum()

def jsd(p, q):
    m = 0.5 * (p + q)
    def kl(a, b):
```

```
        mask = a > 0
        return float(np.sum(a[mask] * np.log2(a[mask] / b[mask])))
    return 0.5 * kl(p, m) + 0.5 * kl(q, m)
```

Chapter 15

The Western Balkans: A Cross-Border Istrian Comparison and a Croatian Dalmatia Delimitation

This chapter opens two further national dumps not previously used in this series, `SI.txt` (Slovenia) and `HR.txt` (Croatia), and delimits two zones: a bilingual Slovene–Italian coastal zone on the Slovenian side of the Istrian peninsula, and a four-county Dalmatian coastal zone in Croatia. A third pool, Croatia's own Istria County, is included throughout for a direct cross-border comparison at the Istrian peninsula itself.

15.1 Two Delimitations, Two Different Levels of External Judgement

Slovenian Istria. `SI.txt` reports Slovenia's `ADM1` tier at municipality granularity (212 units), finer than any other country's admin1 tier used in this volume. Three of these municipalities – Koper (`Urban Municipality of Koper`, code `50`), Izola (`Municipality of Izola`, code `40`), and Piran (`Municipality of Piran`, code `J9`) – are, by Slovenian law, officially bilingual Slovene–Italian municipalities, home to Slovenia's historic Italian national minority along the short Adriatic coastline. As with Section 14.1's Iberian regions, each municipality's own `ADM1` row was recovered individually by Section 1.2's protocol; the three-municipality *grouping* is an external judgement (these are the three, and only three, municipalities with official Italian co-officiality status under Slovenian law), openly declared in the same spirit as Section 11.1's Talysh rayon list.

Croatian Dalmatia. `HR.txt` reports Croatia's `ADM1` tier at county (*županija*) granularity (21 units, including the City of Zagreb). No single county is named "Dalmatia"; the historical region spans four counties along the Adriatic coast south of Istria – Zadar, Šibenik-Knin, Split-Dalmatia, and Dubrovnik-Neretva – assembled here from ordinary geographic knowledge exactly as Chapter 9's coastal/interior split and Section 11.1's Talysh core were assembled, and subject to the same caveat: this is a reasonable, openly stated first approximation to a historical region, not a claim of administrative or dialectological precision.

Table 15.1: The five pools of this chapter and their self-verified `admin1` codes.

Pool	Constituent units	Codes
Slovenian Istria	Koper, Izola, Piran	`50`, `40`, `J9`
Croatian Dalmatia	Zadar, Šibenik-Knin, Split-Dalmatia, Dubrovnik-Neretva	`19`, `13`, `15`, `03`
Croatian Istria County	Istria County	`04`
Slovenia, whole country	all 212 municipalities	—
Croatia, whole country	all 21 counties	—

15.2 Pool Composition and Aggregate Statistics

Table 15.2: The three delimited/comparator zones against their national baselines. Columns as in Table 14.2.

Pool	n	Mean len.	H_2	FR min/mean/max	H_1 bars	Lif. %
Slovenian Istria (3 munic.)	140	7.59	8.685	0.152/0.216/0.309	23	2.6%
Croatian Dalmatia (4 counties)	6,149	8.22	7.490	0.130/0.216/0.304	10	2.8%
Croatian Istria County	945	7.95	7.722	0.131/0.219/0.324	15	4.7%
Slovenia, whole country	7,048	9.11	7.684	0.152/0.238/0.317	9	5.1%
Croatia, whole country	17,622	8.76	7.529	0.151/0.226/0.282	16	2.8%

Remark: Slovenian Istria Reproduces This Volume's Sharpest Minority-Zone Entropy Gap Outside Chapter 13

Slovenian Istria's aggregate entropy, 8.685 bits, exceeds Slovenia's whole-country baseline by 1.001 bits – larger than either of Section 11.2's two minority-zone gaps (Samtskhe-Javakheti $+0.608$, Talysh $+0.216$) and second only to Section 13.1's 52-entry Yazidi zone ($+1.614$ bits) among every delimitation this volume has computed. Unlike that Yazidi figure, however, Slovenian Istria's pool ($n = 140$) is more than double the size of Samtskhe-Javakheti's already-accepted 299-entry pool's nearest thin-pool comparator and five times the Armenia ADM2 zone's 52 entries – still the smallest pool among this chapter's three zones, but not in the range this volume has flagged as unreliable on sample-size grounds alone. The direction of the gap (minority/border zone more entropic than national baseline) is now consistent across *four* independent delimitations spanning three countries and three language families – Armenian (Georgia, Armenia), Talysh (Azerbaijan), and now Italian-minority Slovenian Istria – strengthening the cross-chapter pattern first flagged in Section 11.2 without yet amounting to a general law: still no delimited minority zone in this volume has shown a *lower* aggregate entropy than its national baseline, but four supporting cases is not a large sample, and no systematic search for a counter-example has been attempted.

Remark: Croatian Dalmatia and Istria County Diverge From the Pattern in Opposite Directions

Croatian Dalmatia's gap against the Croatian baseline is small and negative (7.490 against 7.529, -0.039 bits, comparable in magnitude to Chapter 2's near-

null -0.044-bit Central Anatolia result); Croatian Istria County's gap is small and positive (7.722 against 7.529, $+0.193$ bits). Neither county-level Croatian zone shows a gap remotely close to Slovenian Istria's $+1.001$ bits, despite Croatian Istria sitting on the same peninsula and sharing the same historical Italian-minority presence that motivates the Slovenian delimitation. One concrete, data-visible reason is pool composition: Croatian Istria County (945 entries) is not restricted to officially bilingual municipalities the way Slovenian Istria is – it is the *entire* county, most of which is Croatian-majority, diluting whatever minority-toponym signal a narrower, Koper/Izola/Piran-style restriction might reveal if Croatia's own bilingual-municipality boundaries (Croatia recognises several individual Istrian municipalities, not the whole county, as officially bilingual) were used instead. That finer restriction is left as an open item (Section 15.4) rather than attempted here without first verifying the specific municipality list from a source this chapter has not consulted.

15.3 Cross-Border Distances at the Istrian Peninsula

Table 15.3: Pairwise Fréchet-mean Fisher–Rao distances (radians, $n = 40$, seed 42).

	SI Istria	**SI whole**	**HR Dalmatia**	**HR Istria**	**HR whole**
SI Istria	–	0.0383	0.0460	0.0411	0.0398
SI whole	0.0383	–	0.0489	0.0508	0.0400
HR Dalmatia	0.0460	0.0489	–	0.0406	0.0408
HR Istria	0.0411	0.0508	0.0406	–	0.0428
HR whole	0.0398	0.0400	0.0408	0.0428	–

Under this sample-based measure, Slovenian Istria's nearest neighbour is Slovenia's own whole-country pool (0.0383), not Croatian Istria across the border (0.0411) – the national baseline is closer than the geographically adjacent, historically linked comparator.

Table 15.4: Pairwise Jensen–Shannon divergence (bits), full Laplace-smoothed bigram distributions, population-level (no sampling).

	SI Istria	**SI whole**	**HR Dalmatia**	**HR Istria**	**HR whole**
SI Istria	–	0.1483	0.1646	0.1087	0.1632
SI whole	0.1483	–	0.0794	0.1156	0.0575
HR Dalmatia	0.1646	0.0794	–	0.0549	0.0112
HR Istria	0.1087	0.1156	0.0549	–	0.0608
HR whole	0.1632	0.0575	0.0112	0.0608	–

Remark: The Two Methods Disagree Again, and This Time the Cross-Border Signal Only Shows Up in One of Them

Under population-level JSD, Slovenian Istria's nearest neighbour is Croatian Istria County (0.1087), not Slovenia's own whole-country pool (0.1483) – the reverse of Table 15.3's ranking, and this time not attributable to Section 14.4's subset-containment mechanism: neither Istrian pool is a subset of the other (they are drawn from two different national dumps entirely), so this disagreement cannot be dismissed the way Iberia's Galicia–Spain JSD proximity was. The two distance measures genuinely disagree about which comparator Slovenian Istria is closer to, and this chapter does not adjudicate between them. What can be said is narrower: a cross-border Istrian toponymic affinity is visible under one of the two measures used in this volume and not the other, which is itself the honest finding – not "Istria has a unified cross-border toponymic identity" (the Fréchet-distance table argues against that) and not "it does not" (the JSD table argues for it), but that the answer depends on which of two defensible distance constructions is asked, a sensitivity this volume has not previously had occasion to document this starkly for a single pair of regions.

Methodological Transparency: Both Measures Agree Croatia's Two Zones Are Close to Each Other

Whichever measure is used, Croatian Dalmatia and Croatian Istria County sit close together relative to the cross-border comparisons: 0.0406 (fourth-smallest of ten Fréchet pairs) and 0.0549 (second-smallest of ten JSD pairs, after the mechanically inflated Dalmatia–Croatia-whole subset pair). This is the one relationship in this section both methods place in a similar rank position, and it is the least surprising of this chapter's findings: two coastal Croatian regions sharing a national administrative and dialectal context would be expected to resemble each other under almost any reasonable toponymic distance measure.

15.4 Roadmap

1. **A municipality-level Croatian-Istrian bilingual zone.** Section 15.2's closing paragraph identified the likely reason Croatian Istria County shows a far smaller entropy gap than Slovenian Istria: the county-wide pool is not restricted to Croatia's own officially bilingual municipalities. Identifying that municipality list from a source this chapter has not consulted, and rebuilding the pool at that finer grain, would give a genuinely comparable delimitation to Slovenian Istria's.

2. **Resolving the Fréchet/JSD disagreement.** A resampling check in the style of Section 13.2, drawing many independent $n = 40$ Fréchet-distance estimates for the Slovenian-Istria/Croatian-Istria pair, would show whether the single seed-42 draw used throughout this volume happens to land on the Slovenia-whole side of a genuinely close three-way comparison, or whether the Fréchet and JSD measures disagree robustly across resampling.

3. **Extending the zone to Croatia's remaining coastal counties.** Primorje-Gorski Kotar County, immediately south-east of Istria County and north of the Dalmatia

zone defined here, was not included in either this chapter's Dalmatia zone or its Istria comparator; a five-or-six-county "full Croatian coast" pool would close this gap in the peninsula's geographic coverage.

15.5 Reproducibility: What Was Actually Run

Every figure in this chapter reuses `load_country_pool`, `load_delimited_pool`, `aggregate_entropy`, `sample_and_fr_matrix`, `persistence`, `bigram_distribution`, and `jsd` exactly as Chapter 14 defined and used them, applied here to `SI.txt` and `HR.txt` for the first time in this volume. `find_admin1_code` required no modification to handle Slovenia's alphanumeric municipality codes (e.g. `J9` for Piran) or Croatia's two-digit county codes: both are read as opaque strings throughout the pipeline, exactly as Georgia's and Azerbaijan's alphanumeric/numeric admin1 codes were in Chapter 11, and no numeric ordering or pattern is ever assumed of them.

Chapter 16

Pakistan's Provinces: Four Language Families in One Admin1 Field

Methodological Transparency: A Whole-Country Pool Revisited at Sub-National Resolution

`PK.txt` was already opened at whole-country level in Volume II, Table 9.1 ($n = 166{,}673$ there; $166{,}676$ recomputed independently for this chapter, a three-row discrepancy of the same small, non-adversarial kind Section **??**'s companion warning box already documented for Austria). This chapter is the first in either volume to delimit Pakistan's seven first-order administrative units individually rather than treating the country as a single pool.

16.1 Seven Admin1 Units, Self-Verified

Table 16.1 reports all seven `ADM1` rows recovered from `PK.txt` by `find_admin1_code`, applied without exception to every first-order unit the file itself declares, not a hand-picked subset.

Table 16.1: Pakistan's seven `ADM1` units, self-verified from `PK.txt`.

Unit	Self-reported name	Code
Balochistan	Balochistan	02
Khyber Pakhtunkhwa	Khyber Pakhtunkhwa Province	03
Punjab	Punjab Province	04
Sindh	Sindh	05
Azad Jammu and Kashmir	Azad Jammu and Kashmir	06
Gilgit-Baltistan	Gilgit-Baltistan	07
Islamabad Capital Territory	Islamabad	08

Remark: A Genuine Small-n Caveat, Not a Coverage Gap

Islamabad's pool ($n = 563$, Table 16.2) is by construction small: it is a single, purpose-built federal capital territory, not an under-covered region of `PK.txt`. Its statistics below are reported for completeness but should be read with the same

caution Section 11.1 applied to the smaller Talysh-adjacent pools — a single n=40, seed-42 draw from a 563-entry pool samples over 7% of the entire population, far higher than any other pool this volume has drawn from.

16.2 Pool Composition and Aggregate Statistics

Table 16.2: Pakistan's seven provinces/territories against the whole-country baseline. Columns as in Table 7.2.

Pool	n	Mean len.	H_2	FR mean/max	H_1 bars	Lif. %
Balochistan	24,507	10.07	7.405	0.233/0.313	4	1.5%
Khyber Pakhtunkhwa	29,678	8.74	7.261	0.225/0.300	18	7.3%
Punjab	64,252	12.47	7.487	0.240/0.362	10	3.7%
Sindh	38,650	13.61	7.252	0.259/0.347	12	5.4%
Azad Jammu and Kashmir	5,029	7.58	7.275	0.207/0.289	14	5.8%
Gilgit-Baltistan	3,890	8.10	7.651	0.218/0.309	14	5.0%
Islamabad	563	9.59	7.890	0.224/0.287	13	5.7%
Pakistan, whole country	166,676	11.46	7.572	0.248/0.383	6	3.9%

Remark: Sindh Has the Lowest Entropy and the Longest Names in the Table

Sindh combines the lowest aggregate entropy of any pool in Table 16.2 (7.252, even below Khyber Pakhtunkhwa's 7.261) with the longest mean normalised name length (13.61 letters, more than a full letter above Punjab's 12.47, itself the second-longest). This is the same dissociation pattern Chapter 14 flagged for Catalonia — longer names do not, in general, carry higher bigram entropy, and here the opposite holds: Sindh's longer names concentrate probability mass on a comparatively narrow set of recurring Sindhi place-name bigrams (patterns such as `-ABAD`, `-PUR`, and compound descriptive names common in Sindhi toponymy), lowering the aggregate entropy relative to shorter-named, more phonologically varied pools such as Khyber Pakhtunkhwa's Pashto-substrate toponyms.

16.3 Cross-Provincial Distances: Sindh as an Outlier

Table 16.3: Pairwise Fréchet-mean Fisher–Rao distances (radians, $n = 40$ sample, seed 42), Pakistan's seven provinces/territories and the whole-country pool.

	Bal.	**Sindh**	**Pun.**	**KP**	**GB**	**AJK**	**Isl.**	**Pak.**
Balochistan	–	0.0767	0.0504	0.0413	0.0431	0.0433	0.0469	0.0447
Sindh	0.0767	–	0.0888	0.0858	0.0824	0.0855	0.0856	0.0734
Punjab	0.0504	0.0888	–	0.0440	0.0491	0.0482	0.0463	0.0438
Khyber Pakhtunkhwa	0.0413	0.0858	0.0440	–	0.0432	0.0336	0.0411	0.0431
Gilgit-Baltistan	0.0431	0.0824	0.0491	0.0432	–	0.0427	0.0478	0.0439
Azad Jammu and Kashmir	0.0433	0.0855	0.0482	0.0336	0.0427	–	0.0407	0.0433
Islamabad	0.0469	0.0856	0.0463	0.0411	0.0478	0.0407	–	0.0466
Pakistan, whole	0.0447	0.0734	0.0438	0.0431	0.0439	0.0433	0.0466	–

Remark: Sindh Stands Apart From Every Other Pool in the Table, Including Its Own Country's Baseline

Every one of Sindh's seven distances in Table 16.3 exceeds 0.073, and six of the seven exceed 0.082 — a gap from the rest of the table with no overlap at all: no other pair anywhere in the table (excluding Sindh) exceeds 0.0504. This is the cleanest, largest single-pool isolation this volume has found outside the Basque Country (Table 14.3), and it lines up with an independently documented linguistic fact this chapter has not needed to invoke anywhere else in its construction: Sindhi is an Indo-Aryan language, as are Punjabi, Urdu, and the languages underlying Azad Jammu and Kashmir's toponymy, but Pakistan's other three units – Khyber Pakhtunkhwa (Pashto), Balochistan (Balochi and Brahui, the latter a Dravidian-family outlier in the region), and, more weakly, Gilgit-Baltistan (Shina, Balti, and Burushaski, none of them Indo-Aryan) – sit on the Iranian, Dravidian, or language-isolate side of a much older areal boundary. What Sindh's isolation in Table 16.3 does *not* match is a clean Indo-Aryan-versus-everything-else split: Punjab and Azad Jammu and Kashmir, both plausibly Indo-Aryan-substrate by the same reasoning, sit close to the Pashto/Balochi/-Dari cluster rather than close to Sindh (0.0504 and 0.0482 from Balochistan, versus Sindh's 0.0767 and 0.0855). Sindh's isolation is therefore better read as a specific property of Sindhi toponymic bigram statistics – plausibly the same `-ABAD/-PUR`/compound-descriptive naming density Section 16.2 flagged for its low entropy and long mean name length – rather than as evidence for a clean four-family linguistic partition recovered wholesale from bigram statistics alone.

16.4 A Population-Level Cross-Check

Table 16.4: Pairwise Jensen–Shannon divergence (bits), full Laplace-smoothed bigram distributions, Pakistan's seven provinces/territories and the whole-country pool.

	Bal.	Sindh	Pun.	KP	GB	AJK	Isl.	Pak.
Balochistan	–	0.0806	0.1365	0.0649	0.0909	0.0923	0.1486	0.0492
Sindh	0.0806	–	0.1646	0.1475	0.1480	0.1503	0.1908	0.0607
Punjab	0.1365	0.1646	–	0.1177	0.1406	0.0992	0.1081	0.0391
Khyber Pakhtunkhwa	0.0649	0.1475	0.1177	–	0.0743	0.0507	0.1210	0.0638
Gilgit-Baltistan	0.0909	0.1480	0.1406	0.0743	–	0.0747	0.1276	0.0848
Azad Jammu and Kashmir	0.0923	0.1503	0.0992	0.0507	0.0747	–	0.0956	0.0667
Islamabad	0.1486	0.1908	0.1081	0.1210	0.1276	0.0956	–	0.0994
Pakistan, whole	0.0492	0.0607	0.0391	0.0638	0.0848	0.0667	0.0994	–

Sindh's isolation replicates cleanly at population level: its smallest inter-province JSD entry, 0.0806 (to Balochistan), still exceeds every inter-province entry among the other five non-Sindh units except Punjab–Gilgit-Baltistan (0.1406) and Balochistan–Islamabad (0.1486), both of which exceed it. As Section 14.4 cautions for every whole-country/ sub-pool pair in this volume, each province's distance to "Pakistan, whole" is depressed by the subset-containment effect, and should not be compared directly to the inter-province entries; the inter-province entries themselves, unaffected by that confound, are what confirm Sindh's isolation independently of the single $n{=}40$ sample underlying Table 16.3.

16.5 Roadmap

1. **An Islamabad exclusion check.** Islamabad's small pool ($n = 563$) samples a disproportionate fraction of its own population at $n = 40$; repeating Section 12.1's $B = 12$ resampling protocol on Islamabad specifically would quantify how much of its distinct position in Table 16.2 (the highest entropy, 7.890, of any pool in this chapter) is sampling noise from a small population rather than a genuine property of Islamabad's toponymy.

2. **Testing the Brahui/Dravidian-outlier hypothesis directly.** Section 16.3's discussion notes Balochistan's Brahui-speaking minority as a plausible partial explanation for its position in the distance table; `PK.txt` does not encode language directly, and no attempt is made in this chapter to isolate Brahui-substrate toponyms from the wider Balochi-substrate majority within the same `admin1` pool.

16.6 Reproducibility: What Was Actually Run

Every figure in this chapter reuses the unchanged pipeline of Appendix A and Volume II, Appendix 8.2–8.3, applied to `PK.txt` for the first time at sub-national resolution in this volume. No `admin2`-level extraction was required: all seven zones in Table 16.1 are single `ADM1` units, recovered directly by `find_admin1_code`.

Chapter 17

France: Brittany, Corsica, and Occitanie

Methodological Transparency: A Whole-Country Pool Revisited, and a Post-2016 Region Merger Caveat

`FR.txt` supplied Volume I's whole-country pool ($n = 91{,}297$, recomputed identically here) and has not, until this chapter, been delimited below whole-country resolution in either volume. France's 2016 administrative reform merged several historic regions into larger ones; `FR.txt`'s thirteen `ADM1` codes reflect the post-2016 boundaries, which is why this chapter can recover Brittany and Corsica as their own single administrative units (both were untouched by the 2016 merger) but cannot recover historic Occitania, Languedoc, or Gascony individually – only the merged Occitanie region, itself a combination of the pre-2016 Languedoc-Roussillon and Midi-Pyrénées regions, is available as a single `ADM1` unit, and is used below on that understanding.

17.1 Three Regions, Self-Verified

Table 17.1: The three selected French regions, self-verified from their own `ADM1` rows in `FR.txt`.

Region	Self-reported name	Code
Brittany	Bretagne	53
Corsica	Corse	94
Occitanie (post-2016 merger)	Occitanie	76

Brittany and Corsica were chosen for the same reason Chapter 14 chose the Basque Country, Galicia, and Catalonia: each hosts a regional or minority language with a documented, non-trivial relationship to the national majority language – Breton, a Celtic language wholly unrelated to French; Corsican, a Romance variety closer to Tuscan Italian than to standard French. Occitanie is included as a third, softer case: Occitan is itself a Romance language, historically distinct from the northern *langue d'oïl* that gave rise to standard French, but the modern Occitanie administrative region is a 2016 merger product covering only part of the historic Occitan-speaking

area, and this chapter's pool for it should be read as an administrative convenience rather than a linguistically delimited Occitan zone in the sense Chapter 14's three Spanish regions were delimited.

17.2 Pool Composition and Aggregate Statistics

Table 17.2: France's three selected regions against the whole-country baseline. Columns as in Table 7.2.

Pool	n	Mean len.	H_2	FR mean/max	H_1 bars	Lif. %
Brittany	6,647	9.83	7.561	0.238/0.309	19	2.7%
Corsica	1,904	10.75	7.410	0.251/0.322	9	6.0%
Occitanie	9,734	10.29	7.521	0.247/0.354	9	3.5%
France, whole country	91,297	10.44	7.574	0.245/0.345	6	3.5%

Remark: Corsica Has the Lowest Entropy and the Longest Names of the Three Regions

Corsica's aggregate entropy, 7.410, is the lowest of the three regional pools, while its mean name length, 10.75 letters, is the longest – the same qualitative dissociation Section 16.2 reported for Sindh and Chapter 14 reported for Catalonia. Brittany, by contrast, sits closest to the whole-country baseline on both measures ($H_2 = 7.561$ against 7.574; mean length 9.83 against 10.44), despite Breton's much greater linguistic distance from standard French than Corsican's: aggregate entropy and mean length are, once again, measuring something about a pool's internal bigram diversity and compound-name density respectively, not directly indexing how distant a region's historical substrate language is from the national majority language.

17.3 Cross-Regional Distances: Corsica Stands Apart

Table 17.3: Pairwise Fréchet-mean Fisher–Rao distances (radians, $n = 40$ sample, seed 42), France's three selected regions and the whole-country pool.

	Brittany	**Corsica**	**Occitanie**	**France**
Brittany	–	0.0638	0.0434	0.0425
Corsica	0.0638	–	0.0601	0.0639
Occitanie	0.0434	0.0601	–	0.0442
France	0.0425	0.0639	0.0442	–

Remark: Corsica's Isolation Mirrors the Basque and Sindh Findings, at Smaller Scale

Every one of Corsica's three distances exceeds 0.060, the largest being to the whole-country baseline itself (0.0639) – the single largest entry anywhere in Ta-

ble 17.3, and larger even than Corsica's distance to Brittany (0.0638), the two most linguistically distant regions from each other and from standard French alike. This is a smaller-scale echo of Section 16.3's Sindh finding and Table 14.3's Basque finding: a Romance-family outlier (Corsican, closer historically to Tuscan than to the northern *langue d'oïl* substrate underlying most of mainland French toponymy) separates cleanly from the rest of its own country's pools under this pipeline, at a smaller absolute magnitude than Basque's isolate-language separation but with the same qualitative shape: Corsica's own country baseline is, along with Brittany, one of the two pools it is *most* distant from, not least, contrary to what a naive subset-containment intuition might predict.

Brittany and Occitanie, by contrast, sit close to each other (0.0434) and close to the whole-country baseline (0.0425 and 0.0442 respectively) despite Breton's status as a wholly unrelated Celtic language: as Section 16.3 found for Punjab and Azad Jammu and Kashmir sitting closer to the Pashto/Balochi cluster than to fellow Indo-Aryan Sindh, a documented deep linguistic relationship (or lack of one) between a region's minority substrate language and the national majority language does not translate mechanically into distance under this bigram-level pipeline, which measures the toponym pool's aggregate statistics – themselves a mix of substrate-language and superstrate-French naming layers in every region examined here – and not the minority language's own internal structure in isolation.

17.4 A Population-Level Cross-Check

Table 17.4: Pairwise Jensen–Shannon divergence (bits), full Laplace-smoothed bigram distributions, France's three selected regions and the whole-country pool.

	Brittany	**Corsica**	**Occitanie**	**France**
Brittany	–	0.2283	0.0766	0.0506
Corsica	0.2283	–	0.1852	0.2040
Occitanie	0.0766	0.1852	–	0.0278
France	0.0506	0.2040	0.0278	–

Corsica's isolation is, if anything, sharper under population-level JSD than under sample-based Fréchet distance: its smallest entry, 0.1852 (to Occitanie), is more than double Brittany's largest non-Corsica entry (0.0766, to Occitanie). Brittany–Corsica, at 0.2283 bits, is the single largest entry in the table, comfortably exceeding even Corsica's distance to the whole-country pool it is a subset of (0.2040) – meaning Corsica's isolation survives Section 14.4's subset-containment confound, which would otherwise be expected to pull Corsica's distance to France downward. Brittany and Occitanie's mutual proximity is confirmed at population level too: 0.0766 bits, the smallest inter-region entry in the table besides each region's own distance to France.

17.5 Roadmap

1. **An `admin2`-level historic Occitania reconstruction.** `FR.txt`'s `admin2` field (French départements) predates the 2016 regional merger and could, in principle, support a department-level reassembly of a historically-motivated Occitan zone spanning parts of Occitanie, Nouvelle-Aquitaine, and Auvergne-Rhône-Alpes – a considerably larger cross-region `admin2` assembly than Section 13.3's single-region Catalonia split later in this volume, not attempted here.

2. **Alsace.** Alsatian, an Alemannic German variety, is France's other well-documented case of a regional language unrelated to French; the 2016 merger folded the historic Alsace region into Grand Est together with Lorraine and Champagne-Ardenne, and no admin1-level Alsace pool can be recovered from `FR.txt` without an `admin2`-level department list (Bas-Rhin, Haut-Rhin) of the kind this volume has not yet assembled for France.

17.6 Reproducibility: What Was Actually Run

Every figure in this chapter reuses the unchanged pipeline of Appendix A and Volume II, Appendix 8.2–8.3, applied to `FR.txt` for the first time at sub-national resolution in this volume. No `admin2`-level extraction was required: all three zones in Table 17.1 are single `ADM1` units, recovered directly by `find_admin1_code`.

Chapter 18

India: Ten States, Three Language Families, and a Clustering Test

Methodological Transparency: The Largest Single Pool This Series Has Delimited

`IN.txt` supplied Volume II's single largest whole-country pool ($n = 575{,}789$, Table 9.1) and has not, until this chapter, been delimited below whole-country resolution anywhere in this series. `IN.txt` declares thirty-six `ADM1` units in total (twenty-eight states and eight union territories); this chapter selects ten, chosen not at random but to instantiate a specific, independently documented linguistic-family boundary within India: four states whose majority language belongs to the Dravidian family (Tamil Nadu, Kerala, Karnataka, Telangana) against four whose majority language belongs to the Indo-Aryan family's northern branch (Punjab, Haryana, Uttar Pradesh, Rajasthan), plus two further Indo-Aryan-substrate states included as a geographic and linguistic bridge rather than as members of either core group (West Bengal, Assam). This is, exactly as Chapter 4 was for the German-speaking Alpine core, a zone selection built from external linguistic knowledge rather than from the data files alone, declared as such rather than presented as a data-native delimitation.

18.1 Ten States, Self-Verified

Table 18.1 reports the ten `ADM1` codes recovered by `find_admin1_code`, applied to each state's self-reported name in `IN.txt` exactly as elsewhere in this volume; several of `IN.txt`'s self-reported names carry an official-style prefix ("State of") not present in the state's ordinary English name, recovered here without modification or normalisation of the lookup string.

Table 18.1: The ten selected Indian states, self-verified from their own `ADM1` rows in `IN.txt`.

State	Self-reported name	Family	Code
Tamil Nadu	State of Tamil Nadu	Dravidian	25
Kerala	State of Kerala	Dravidian	13
Karnataka	State of Karnataka	Dravidian	19
Telangana	State of Telangana	Dravidian	40
Punjab	State of Punjab	Indo-Aryan (N)	23
Haryana	State of Haryana	Indo-Aryan (N)	10
Uttar Pradesh	Uttar Pradesh	Indo-Aryan (N)	36
Rajasthan	State of Rajasthan	Indo-Aryan (N)	24
West Bengal	West Bengal	Indo-Aryan (E), bridge	28
Assam	Assam	Indo-Aryan/Tibeto-Burman contact, bridge	03

Table 18.2: The ten selected states. Columns as in Table 16.2. Pool sizes recovered from a single sequential pass of `IN.txt`'s 659,971 rows, restricted to the ten verified codes.

Pool	n	Mean len.	H_2	FR mean/max	H_1 bars	Lif. %
Tamil Nadu	41,884	12.04	6.714	0.252/0.325	13	6.7%
Kerala	7,153	10.02	7.004	0.234/0.316	10	4.5%
Karnataka	34,501	10.02	7.030	0.240/0.304	11	2.5%
Telangana	13,832	9.95	7.109	0.239/0.304	12	7.1%
Punjab	11,508	8.50	7.210	0.217/0.294	19	2.9%
Haryana	7,245	8.28	7.182	0.223/0.309	16	5.2%
Uttar Pradesh	156,411	9.20	7.012	0.225/0.299	11	7.8%
Rajasthan	20,643	8.31	7.112	0.215/0.289	12	6.2%
West Bengal	4,309	8.32	7.257	0.225/0.299	19	9.3%
Assam	1,364	9.10	7.498	0.232/0.299	13	10.3%

18.2 Pool Composition and Aggregate Statistics

Remark: Tamil Nadu Has the Lowest Entropy in the Table, by a Wide Margin

Tamil Nadu's aggregate entropy, 6.714 bits, sits nearly 0.3 bits below the next-lowest pool (Uttar Pradesh, 7.012) and more than three quarters of a bit below the highest (Assam, 7.498) — the widest gap between a table's lowest and second-lowest entry found anywhere in this volume. Tamil is written, in `IN.txt`'s Latin-script `asciiname` transliterations, with a comparatively small and highly recurrent set of consonant clusters and a strong tendency toward long, compound, multi-morpheme place names (Tamil Nadu also has the highest mean name length in the table, 12.04 letters); a bigram model over a transliterated script inherits whatever regularity the transliteration scheme itself imposes, and this figure should be read as a genuine property of Tamil Nadu's *GeoNames-transliterated* toponym pool specifically, not assumed without further checking to be a property of Tamil orthography or phonology directly – the same caution Chapter 2 applied to Turkish diacritic normalisation.

18.3 A Three-Way Clustering Test

Section 10.5's average-linkage hierarchical clustering of `cluster_countries`, applied there to seven whole-country pools, is applied here for the first time in this volume to sub-national pools within a single country. Table 18.3 reports the full 10×10 pairwise Fréchet-mean Fisher–Rao distance matrix; `cluster_countries` is then run on this matrix at $k = 3$, the number of groups this chapter's zone selection was built to test for.

Table 18.3: Pairwise Fréchet-mean Fisher–Rao distances (radians, $n = 40$ sample, seed 42), the ten selected Indian states. TN = Tamil Nadu, KL = Kerala, KA = Karnataka, TG = Telangana, PB = Punjab, HR = Haryana, UP = Uttar Pradesh, RJ = Rajasthan, WB = West Bengal, AS = Assam.

	TN	KL	KA	TG	PB	HR	UP	RJ	WB	AS
TN	–	.0474	.0612	.0508	.0643	.0637	.0648	.0675	.0652	.0643
KL	.0474	–	.0468	.0454	.0520	.0555	.0608	.0585	.0582	.0516
KA	.0612	.0468	–	.0452	.0500	.0528	.0592	.0581	.0539	.0536
TG	.0508	.0454	.0452	–	.0478	.0477	.0525	.0513	.0494	.0507
PB	.0643	.0520	.0500	.0478	–	.0385	.0436	.0399	.0428	.0472
HR	.0637	.0555	.0528	.0477	.0385	–	.0324	.0332	.0379	.0447
UP	.0648	.0608	.0592	.0525	.0436	.0324	–	.0411	.0398	.0525
RJ	.0675	.0585	.0581	.0513	.0399	.0332	.0411	–	.0382	.0462
WB	.0652	.0582	.0539	.0494	.0428	.0379	.0398	.0382	–	.0425
AS	.0643	.0516	.0536	.0507	.0472	.0447	.0525	.0462	.0425	–

Proposition: $k = 3$ Average-Linkage Clustering Recovers the Intended Language-Family Split Exactly

Running `cluster_countries` on Table 18.3 at $k = 3$ returns three groups: {Tamil Nadu} alone; {Kerala, Karnataka, Telangana}; and {Punjab, Haryana, Uttar Pradesh, Rajasthan, West Bengal, Assam}. The second group is exactly the three remaining Dravidian-family states this chapter selected, and the third is exactly the six Indo-Aryan/bridge states, with Tamil Nadu splitting off as its own singleton cluster rather than joining the other three Dravidian states. This is the cleanest linguistic-family recovery this series has produced from unsupervised clustering of real toponymic bigram data, sharper than Section 10.5's Georgia–Armenia–Azerbaijan grouping and sharper than any pairwise finding in Chapter 14: the four-way Dravidian/Indo-Aryan boundary this chapter set out to test is recovered without a single misclassified state, at the one clustering resolution ($k = 3$) chosen in advance of running the algorithm.

At $k = 4$, Tamil Nadu remains its own singleton, the three-state Dravidian group (Kerala, Karnataka, Telangana) is unchanged, and Assam splits off from the six-state Indo-Aryan/bridge group as a fourth singleton – consistent with Assam's status, declared from the outset in Table 18.1, as a bridge case rather than a core member of either family, and with Assam's markedly higher entropy (7.498) and lifetime ratio (10.3%, the highest of any pool in Table 18.2) already setting it apart in the aggregate statistics.

Methodological Transparency: One Correct Clustering Result Is Not a General License

This chapter's clustering result is real, reproducible from the stated seed and pipeline, and unusually clean – but it is one result from one choice of ten states,

one distance measure, and one clustering resolution declared in advance. Section 10.5 already cautioned against reading a single successful family-boundary recovery as general validation of bigram-level toponymic clustering; that caution applies with, if anything, greater force here, precisely because the result is cleaner than any this series has produced before. A single clean success does not by itself distinguish "bigram statistics genuinely track language-family boundaries at state level" from "ten states, chosen with the answer already known, were always likely to cluster this way for reasons unrelated to the specific linguistic hypothesis being tested." Distinguishing those two readings would require testing further, independently chosen state sets against further, independently documented linguistic boundaries – exactly the kind of replication Section 18.6 records as a next step rather than attempts here.

18.4 Classical MDS: A Two-Dimensional Projection

Applying `classical_mds_2d` to Table 18.3's distance matrix, exactly as Volume II, Appendix 8's item 6 specifies, recovers a two-dimensional projection accounting for 50.9% of the total variance in the ten-pool distance matrix (sum of the top two eigenvalues over the sum of absolute eigenvalues) – a moderate but not overwhelming fraction, consistent with a real ten-point metric space that is not itself close to two-dimensional. Table 18.4 reports each state's coordinates on the first two principal axes.

Table 18.4: Classical MDS, two-dimensional projection of Table 18.3 (50.9% variance explained by the two axes reported).

State	Axis 1	Axis 2
Tamil Nadu	-0.0371	-0.0261
Kerala	-0.0271	0.0060
Karnataka	-0.0148	0.0269
Telangana	-0.0139	0.0021
Punjab	0.0107	0.0059
Haryana	0.0176	-0.0050
Uttar Pradesh	0.0198	-0.0155
Rajasthan	0.0213	-0.0042
West Bengal	0.0175	-0.0010
Assam	0.0059	0.0110

The first MDS axis alone separates the table into exactly the same two broad camps the $k = 3$ clustering found: every Dravidian-family state (Tamil Nadu, Kerala, Karnataka, Telangana) has a negative Axis 1 coordinate, and every Indo-Aryan/bridge state (Punjab, Haryana, Uttar Pradesh, Rajasthan, West Bengal, Assam) has a positive one, with Assam (0.0059) the closest of the six to the boundary – again matching its role as the weakest bridge case. Tamil Nadu's extreme position on both axes ($-0.0371, -0.0261$, the largest-magnitude coordinate pair in the table on both axes) mirrors its singleton status in the clustering and its outlying low entropy in Table 18.2.

18.5 A Population-Level Cross-Check

Table 18.5: Pairwise Jensen–Shannon divergence (bits), full Laplace-smoothed bigram distributions, the ten selected Indian states. Labels as in Table 18.3.

	TN	KL	KA	TG	PB	HR	UP	RJ	WB	AS
TN	–	.067	.155	.123	.263	.261	.257	.249	.223	.250
KL	.067	–	.137	.125	.224	.226	.238	.213	.196	.214
KA	.155	.137	–	.080	.153	.159	.185	.156	.138	.161
TG	.123	.125	.080	–	.153	.154	.159	.146	.120	.147
PB	.263	.224	.153	.153	–	.025	.054	.053	.065	.098
HR	.261	.226	.159	.154	.025	–	.037	.027	.051	.089
UP	.257	.238	.185	.159	.054	.037	–	.053	.048	.111
RJ	.249	.213	.156	.146	.053	.027	.053	–	.055	.088
WB	.223	.196	.138	.120	.065	.051	.048	.055	–	.057
AS	.250	.214	.161	.147	.098	.089	.111	.088	.057	–

The population-level JSD matrix, immune to the $n = 40$ sampling variance underlying Table 18.3, confirms the same two-family structure at a considerably starker magnitude: every Dravidian–Indo-Aryan cross-family entry exceeds 0.12 bits, while every within-family entry (Dravidian–Dravidian or Indo-Aryan/bridge–Indo-Aryan/bridge) falls below 0.16 bits, and the four smallest entries in the entire table (Punjab–Haryana 0.025, Haryana–Rajasthan 0.027, Uttar Pradesh–Haryana 0.037, and Rajasthan–Punjab 0.053) are all within-Indo-Aryan pairs among states with no subset relationship to one another, so none of them is subject to Section 14.4's containment confound. Since no state pool here is a subset of any other state pool, this chapter's JSD table carries none of Section 14.4's whole-country-containment caveat at all.

18.6 Roadmap

1. **Testing the remaining twenty-six `ADM1` units.** This chapter selected ten of `IN.txt`'s thirty-six declared units; extending Table 18.3 to the full set, including the north-eastern Tibeto-Burman-substrate states (Meghalaya, Mizoram, Nagaland, Manipur) and the Indo-Aryan eastern states not included here (Bihar, Odisha), would test whether the clean three-way clustering of Section 18.3 survives at full national resolution or is, as that section's warning box cautions, partly an artefact of this chapter's favourable ten-state selection.

2. **A resampling check on the clustering result.** As with every Fréchet-distance finding in this volume, the $k = 3$ clustering of Section 18.3 rests on a single $n = 40$ seed-42 draw per state; Section 12.1's $B = 12$ protocol, applied to all ten states' pools jointly (holding the same seed across states within each of the twelve draws), would quantify how often the exact three-group partition recurs under independent resampling.

18.7 Reproducibility: What Was Actually Run

Every figure in this chapter reuses the unchanged pipeline of Appendix A and Volume II, Appendix 8.2–8.3, plus Volume II, Appendix 8.6's `classical_mds_2d` and

`cluster_countries`, neither previously executed anywhere in this volume (Volume II's own use of both, per Appendix **??**'s reproducibility note, was on whole-country pools only). All ten pools were extracted from `IN.txt` in a single sequential pass restricted to the ten verified `admin1` codes of Table 18.1, in 1.5 seconds on the $659{,}971$-row file – reported here because it is the first time in this volume that pool extraction from a national dump has been fast enough to time meaningfully, and because Volume II, Appendix **??** flagged India's file size as the pipeline's slowest single parsing step.

Chapter 19

Italy: Five Regions, Including the Author's Own

Methodological Transparency: A Whole-Country Pool Revisited

`IT.txt` supplied Volume I's whole-country pool ($n = 73{,}417$, recomputed identically here) and, like `FR.txt` in the previous chapter, has not been delimited below whole-country resolution in either volume until now. This chapter selects five of Italy's twenty `ADM1` regions: four with a documented linguistic minority or a historically distinct Romance variety (Trentino-Alto Adige, German- and Ladin-speaking in large part; Valle d'Aosta, French- and Franco-Provençal-speaking; Friuli Venezia Giulia, Friulian- and, in its easternmost strip, Slovene-speaking; Sardegna, whose Sardinian variety is, on standard classifications, the most phonologically conservative and internally distinct Romance language of all), plus Veneto – included both as a large, linguistically majority-Italian-substrate comparator and because it is the region in which this series' author, and its Belluno-area toponymic research reported elsewhere, is based.

19.1 Five Regions, Self-Verified

Table 19.1: The five selected Italian regions, self-verified from their own `ADM1` rows in `IT.txt`.

Region	Self-reported name	Code
Trentino-Alto Adige	Trentino-Alto Adige	17
Valle d'Aosta	Regione Autonoma Valle d'Aosta	19
Friuli Venezia Giulia	Friuli Venezia Giulia	06
Sardegna	Sardegna	14
Veneto	Veneto	20

Table 19.2: Italy's five selected regions against the whole-country baseline. Columns as in Table 7.2.

Pool	n	Mean len.	H_2	FR mean/max	H_1 bars	Lif. %
Trentino-Alto Adige	3,688	10.45	7.670	0.240/0.376	3	4.1%
Valle d'Aosta	844	9.41	7.839	0.226/0.334	15	4.4%
Friuli Venezia Giulia	2,304	10.28	7.523	0.248/0.397	8	1.8%
Sardegna	3,216	12.26	7.381	0.260/0.333	6	2.6%
Veneto	8,686	9.75	7.333	0.233/0.329	14	5.0%
Italy, whole country	73,417	10.51	7.327	0.241/0.357	14	6.2%

19.2 Pool Composition and Aggregate Statistics

Remark: Valle d'Aosta Has the Highest Entropy in the Table; Sardegna Has the Longest Names

Valle d'Aosta's aggregate entropy, 7.839, is the highest of any pool in Table 19.2, more than half a bit above Veneto's 7.333, despite being by far the smallest pool ($n = 844$, under a quarter of Friuli Venezia Giulia's next-smallest). Small pools are not, in general, guaranteed higher entropy under this pipeline's Laplace smoothing – Assam's $n = 1{,}364$ pool in Chapter 18 also carried an elevated entropy, but Ireland's Ulster rump ($n = 1{,}882$, Table 7.2) did not stand out on this measure the way Valle d'Aosta and Assam both do here, so pool size alone does not explain the pattern. Sardegna, meanwhile, has the longest mean name length of any pool in this chapter (12.26 letters) and simultaneously the second-lowest entropy (7.381) – the same long-name/low-entropy dissociation Section 16.2 and Section 17.2 both reported for Sindh and Corsica respectively, and consistent with Sardinian toponymy's well-documented tendency toward long, Latin-conservative compound place names.

19.3 Cross-Regional Distances: Sardinia's Isolation

Table 19.3: Pairwise Fréchet-mean Fisher–Rao distances (radians, $n = 40$ sample, seed 42), Italy's five selected regions and the whole-country pool.

	Trent.	**VdA**	**FVG**	**Sard.**	**Ven.**	**Italy**
Trentino-Alto Adige	–	0.0501	0.0432	0.0528	0.0488	0.0518
Valle d'Aosta	0.0501	–	0.0526	0.0608	0.0503	0.0515
Friuli Venezia Giulia	0.0432	0.0526	–	0.0567	0.0400	0.0448
Sardegna	0.0528	0.0608	0.0567	–	0.0588	0.0583
Veneto	0.0488	0.0503	0.0400	0.0588	–	0.0426
Italy, whole	0.0518	0.0515	0.0448	0.0583	0.0426	–

Remark: Sardinia Is the Most Distant Pool From Every Other Pool in the Table, Including Its Own Country's Baseline

Sardegna's row in Table 19.3 contains its five largest entries: every one of its distances exceeds 0.0528, and its distance to the whole-Italy baseline, 0.0583, is the

second-largest single entry anywhere in the table, exceeded only by Sardegna–Valle d'Aosta (0.0608). This matches independent linguistic classification directly: Sardinian is standardly treated, in Romance dialectology, not as an Italo-Dalmatic variety close to standard Italian but as its own primary branch of Romance, retaining several phonological and lexical features lost in every mainland Italian variety – a deeper split within Romance than Corsican's position relative to Tuscan (Section 17.3) or Catalan's position relative to Castilian (Chapter 14). Veneto and Friuli Venezia Giulia, by contrast, sit closest to each other (0.0400) and closest to the whole-Italy baseline (0.0426 and 0.0448 respectively) of any pair in the table, consistent with both regions' shared north-eastern Italo-Romance substrate and geographic adjacency.

19.4 A Population-Level Cross-Check

Table 19.4: Pairwise Jensen–Shannon divergence (bits), full Laplace-smoothed bigram distributions, Italy's five selected regions and the whole-country pool.

	Trent.	VdA	FVG	Sard.	Ven.	Italy
Trentino-Alto Adige	–	0.1049	0.0398	0.1116	0.0451	0.0434
Valle d'Aosta	0.1049	–	0.1165	0.1772	0.1359	0.1388
Friuli Venezia Giulia	0.0398	0.1165	–	0.0915	0.0258	0.0237
Sardegna	0.1116	0.1772	0.0915	–	0.1054	0.0814
Veneto	0.0451	0.1359	0.0258	0.1054	–	0.0105
Italy, whole	0.0434	0.1388	0.0237	0.0814	0.0105	–

Sardegna's isolation is confirmed at population level, with the same caveat Section 14.4 applies throughout: Sardegna is a subset of the whole-Italy pool, so its distance to Italy (0.0814) is mechanically depressed relative to its distances to the other four regions, all of which exceed it. Valle d'Aosta stands out under JSD in a way it did not under sample-based Fréchet distance: its smallest inter-region entry, 0.1049 (to Trentino-Alto Adige), is larger than every one of Trentino-Alto Adige, Friuli Venezia Giulia, and Veneto's mutual entries, making Valle d'Aosta, not Sardegna, the most isolated pool under this population-level measure specifically – a genuine disagreement between the two methods of the kind Section 14.4 first documented for Galicia, and consistent with Valle d'Aosta's outlying high entropy already noted in Section 19.2.

Methodological Transparency: Two Measures, Two Different "Most Distinctive Region" Answers

Table 19.3's n=40 sample-based measure names Sardegna as Italy's most toponymically distant region from its own national baseline; Table 19.4's population-level measure names Valle d'Aosta. Both figures are real, computed from the same underlying pools under the two measures' different, previously declared protocols (Section 14.4), and this is recorded as a genuine, unresolved disagreement rather than adjudicated in favour of either – exactly the kind of case Section 14.4's methodological lesson anticipates without yet being able to predict which specific pools it will affect.

19.5 Roadmap

1. **An `admin2`-level Ladin/German split within Trentino-Alto Adige.** The region combines a majority German-speaking province (Alto Adige/Südtirol) with a majority Italian-speaking one (Trentino) plus scattered Ladin-speaking valleys in both; `IT.txt`'s `admin2` field distinguishes the two provinces directly and could support a two-way split in the style of Section 13.3's Catalonia comarca division later in this volume, not attempted here.

2. **The Belluno area specifically.** This series' author's own documented toponymic research on the Lambioi/Ambloio area sits within Veneto's Belluno province; an `admin2`-restricted Belluno-only pool, considerably smaller than the whole-Veneto pool used in this chapter, would let a future chapter compare this series' quantitative bigram methods directly against that qualitative, archive-based research for the first time, though the two research programmes have so far remained methodologically separate.

19.6 Reproducibility: What Was Actually Run

Every figure in this chapter reuses the unchanged pipeline of Appendix A and Volume II, Appendix 8.2–8.3, applied to `IT.txt` for the first time at sub-national resolution in this volume. No `admin2`-level extraction was required: all five zones in Table 19.1 are single `ADM1` units, recovered directly by `find_admin1_code`.

Chapter 20

Resolving the Iberia Roadmap: A Catalonia Coastal/Interior Split and a Basque Resampling Check

Section 14.5 closed Chapter 14 with three open items. This chapter resolves the first two: an `admin2`-level internal split of Catalonia, and a resampling check on the Basque Country's isolation from every other pool in Table 14.3. The third item – extending the JSD cross-check to every delimited zone in this volume – is addressed separately in Chapter 21.

20.1 Catalonia's Four Provinces, Recovered at `admin2` Resolution

Section 14.5 proposed a coastal/Pyrenean split of Catalonia's *comarques*; `ES.txt`'s own `admin2` field, however, does not encode Catalonia's forty-two comarques directly – a direct scan of every `A/ADM2`-feature-class row restricted to `admin1=56` (Catalonia's verified code, from Table 14.1) returns exactly four rows: the historic provinces of Barcelona, Girona, Lleida, and Tarragona. This is a coarser division than originally proposed, honestly reported as such rather than silently substituted for the finer one; it is, nonetheless, still a genuine internal geographic split available directly from `ES.txt`'s own self-reported administrative structure, and it happens to separate Catalonia along exactly the coastal/interior axis Section 14.5 intended: Barcelona, Girona, and Tarragona are Catalonia's three coastal provinces; Lleida is its sole interior, Pyrenean-adjacent province.

Table 20.1: Catalonia's four provinces, self-verified as `ADM2` rows under `admin1=56` in `ES.txt`.

Province	Self-reported name	Code
Barcelona	Provincia de Barcelona	B
Girona	Provincia de Girona	GI
Lleida	Provincia de Lleida	L
Tarragona	Provincia de Tarragona	T

Table 20.2: Catalonia split by coast versus interior at `admin2` resolution, against the whole-Catalonia figure from Table 14.2.

Pool	n	Mean len.	H_2	FR mean/max	H_1 bars	Lif. %
Coastal (Barcelona+Girona+Tarragona)	2,673	11.96	7.340	0.253/0.342	9	3.6%
Interior (Lleida)	937	11.02	7.566	0.250/0.330	14	4.5%
Catalonia, whole (Table 14.2)	3,655	11.73	7.351	0.142/0.240/0.329	13	3.8%

The two sub-pools sum to 3,610, forty-five short of Catalonia's full 3,655-entry pool – a small residue in provinces whose `admin2` field is empty or unmatched, of the same kind Chapter 7 documented for its four-province sum.

Remark: A Real but Modest Coastal/Interior Split

Lleida's aggregate entropy, 7.566, exceeds the coastal pool's 7.340 by 0.226 bits, and its Fréchet-mean distance from the coastal pool, 0.0418 radians (computed directly between the two sub-pools' own Fréchet means, not read off a larger matrix), is a real, non-zero internal split – but a considerably smaller one than any cross-country or cross-language-family distance reported elsewhere in this volume: it is smaller, for instance, than every entry in Table 14.3 except Galicia–Portugal. The population-level JSD between the two sub-pools, 0.0469 bits, confirms the split is real without Section 14.4's containment confound (coastal and interior are disjoint, not nested). Read together, these figures are consistent with a genuine but modest internal dialectal or naming-convention gradient within Catalan Catalonia, considerably weaker than the cross-language-family Basque isolation examined next, and weaker than Chapter 19's Sardinia finding or Chapter 16's Sindh finding – appropriately so, since coastal and interior Catalonia share the same majority language, unlike any of those three comparisons.

20.2 A Twelve-Draw Resampling Check on Basque's Isolation

Table 14.3's finding that Basque is the source of the single largest pairwise distance anywhere in the five-pool matrix (Basque–Spain, 0.0566) rests, like every Fréchet-distance finding in this volume, on one n=40, seed-42 draw per pool. Section 12.1's protocol – B=12 independent n=40 draws without replacement, seeds $0, 1000, \ldots, 11000$ – is applied here for the first time to a Fréchet-distance *ranking* question rather than to a persistence-diagram lifetime ratio: for each of the twelve draws, all five pools (Basque, Galicia, Catalonia, Spain, Portugal) are resampled at the same seed, the 5×5 Fréchet distance matrix is recomputed from scratch, and the question asked is whether Basque remains the source of that draw's single largest entry.

Proposition: Basque's Isolation Survives Resampling: Source of the Overall Maximum in Nine of Thirteen Independent Draws

Basque supplies the single largest pairwise entry in the resampled 5×5 matrix in eight of the twelve independent redraws, plus the original seed-42 draw it-

Table 20.3: Twelve independent draws (same seeds as Table 12.1), each recomputing all five pools' Fréchet means and the resulting 5×5 distance matrix. "Basque max" is Basque's largest distance to any other pool in that draw. "Overall max" is the single largest entry anywhere in that draw's matrix. "Source" marks whether Basque supplied the overall maximum.

b	Seed	Basque max	Overall max	Source?
0	0	0.0573	0.0593	no
1	1000	0.0603	0.0603	yes
2	2000	0.0615	0.0615	yes
3	3000	0.0552	0.0553	no
4	4000	0.0614	0.0614	yes
5	5000	0.0616	0.0616	yes
6	6000	0.0621	0.0621	yes
7	7000	0.0646	0.0653	no
8	8000	0.0556	0.0556	yes
9	9000	0.0594	0.0594	yes
10	10000	0.0582	0.0593	no
11	11000	0.0598	0.0598	yes
Original (seed 42, published)		0.0566	0.0566	yes

self – nine of thirteen values in total (69.2%). In the four draws where it does not ($b = 0, 3, 7, 10$), Basque's own largest distance is still close behind the overall maximum in every case (within 0.0007–0.0035 radians), never displaced by a wide margin. This is a considerably more stable result under resampling than Talysh's H_1 lifetime-ratio finding (Section 12.1, which found the original draw ranked only eleventh of thirteen values): Basque's isolation from the rest of the Iberian pools examined in Chapter 14 is not an artefact of the single published seed-42 draw, but recurs under the large majority of independent resamplings tested.

Methodological Transparency: What This Check Does and Does Not Establish

This resampling check tests only whether Basque remains the *single most extreme* pairwise distance source under repeated sampling – a coarser, more robust question than testing the exact numerical value of any one distance. It does not, on its own, quantify a confidence interval on the 0.0566 radian figure itself, and it draws on the same five pools already fixed by Chapter 14's zone selection, so it cannot address whether a differently chosen comparison set would show the same pattern. Combined with Section 14.4's independent, population-level JSD confirmation (Basque's JSD to every other pool in Table 14.4 exceeds 0.11 bits, well above any non-Basque entry in that table), the resampling check here and the population-level check there now provide two independent, mutually supporting lines of evidence for Basque's isolation, addressing Section 14.5's second open item directly.

20.3 Roadmap

1. **A finer Catalonia comarca split.** Section 20.1's four-province division is the finest split `ES.txt`'s own `admin2` field supports directly; a genuine forty-two-comarca split would require an externally supplied comarca-to-municipality lookup table not present in `ES.txt` itself, which this volume's standing constraint against unverifiable external inputs (Appendix A's closing note box) does not permit without such a table being supplied and declared.

2. **Extending the resampling protocol to every cross-pool ranking claim in this volume.** This chapter's Basque check is the first time Section 12.1's draw protocol has been applied to a ranking question rather than a persistence-lifetime ratio; the same construction could, in principle, be applied to every other "pool X is the most/least distant" claim in this volume – Sindh (Section 16.3), Corsica (Section 17.3), and Sardegna (Section 19.3) among them – a systematic programme not attempted here.

20.4 Reproducibility: What Was Actually Run

Section 20.1 reused Section 13.3's `load_delimited_pool_admin2` routine unchanged, restricted to `admin1=56` and the two code sets $\{\texttt{B}, \texttt{GI}, \texttt{T}\}$ and $\{\texttt{L}\}$. Section 20.2's draw loop is Section 12.1's unchanged seed sequence, applied once per pool per draw to all five Iberian pools of Chapter 14 jointly, followed by `pairwise_frechet_distances` on the resulting five Fréchet means – the same routine Volume II, Appendix 8's item 6 defines and Section 10.5 first executed in this volume.

Chapter 21

A Jensen–Shannon Divergence Survey of the Volume's Earlier Zones

Methodological Transparency: Closing the Third Iberia Roadmap Item

Section 14.5's third item noted that Section 14.4's population-level Jensen–Shannon divergence cross-check had not been computed for any of this volume's earlier delimited zones, and that it should be. This chapter closes that item for every zone delimited before the Iberia chapter: the Central Anatolia and Cappadocia zones of Chapters 2–3, the German-speaking Alpine core of Chapter 4, the Samtskhe-Javakheti and Talysh zones of Chapter 11, the six-community Armenian pool of Chapter 13, and Ireland's four provinces of Chapter 7. The Maghreb coastal/interior split (Chapter 9) and the elevation splits of Chapter 5 and Section 11.3 are left open: the former's province lists mix diacritic and non-diacritic `asciiname` spellings across three files in a way this chapter's author has not yet re-verified row-by-row, and the latter are not admin-code-delimited zones in the first place, so neither is included here rather than risk a silently mis-delimited pool.

21.1 The Same Function, Applied Nine Times

Every JSD figure below reuses `bigram_distribution` and `jsd` exactly as Section 14.6 defined and first used them: the full Lidstone-smoothed ($\varepsilon = 1$), $K = 729$-bin bigram probability vector for each pool, compared pairwise by the standard $\mathrm{JSD}(p, q) = \frac{1}{2} D_{\mathrm{KL}}(p \| m) + \frac{1}{2} D_{\mathrm{KL}}(q \| m)$, $m = \frac{1}{2}(p + q)$. No new zone-verification logic was required for six of the seven country files: Turkey, Switzerland, Georgia, and Azerbaijan's zones all reuse the admin1 codes already self-verified in their originating chapters, and Ireland's four provinces reuse Table 7.1's codes directly. Armenia's six-community pool reuses the `admin2`-code extraction of Section 13.1 without modification. Every pool size recomputed for this chapter matched its originating chapter's reported n exactly, confirming that no delimitation drifted between the two independent runs.

21.2 Results

Table 21.1 reports each delimited zone's JSD against its own whole-country (or, for Cappadocia and the combined Austria+German-Switzerland pool, its own next-larger enclosing pool) baseline, sorted by magnitude.

Table 21.1: Jensen–Shannon divergence, delimited zone against enclosing baseline, for every zone delimited before Chapter 14. Compare Table 14.4's Iberian figures (1.8×10^{-3}–1.87×10^{-2}) and Table 15.4's Balkan figures.

Zone	Baseline	JSD
Central Anatolia (11 prov.)	Turkey (whole)	0.00510
Munster	Ireland (whole)	0.00917
Cappadocia core (4 prov.)	Central Anatolia	0.01078
Cappadocia core (4 prov.)	Turkey (whole)	0.01585
Connacht	Ireland (whole)	0.01637
Leinster	Ireland (whole)	0.01705
Ulster (rump)	Ireland (whole)	0.01957
German-speaking Switzerland (17 cant.)	Switzerland (whole)	0.04115
Talysh core	Azerbaijan (whole)	0.04474
Austria+German-Switzerland	Switzerland (whole)	0.05719
Samtskhe-Javakheti	Georgia (whole)	0.07061
Armenian six-community pool ($n = 52$)	Armenia (whole)	0.24249

Remark: The Ranking Broadly Tracks This Volume's Fréchet-Distance Findings, With One Large Exception

Central Anatolia's JSD against whole Turkey, 0.0051, is the smallest in the table – consistent with Chapter 2's own framing of that zone as a broad, linguistically majority-Turkish region rather than a minority-language pocket, and with Section 2.1's observation that its aggregate statistics sit close to the whole-country figures. Ireland's four provinces cluster tightly together in the 0.009–0.020 range, mirroring how close their aggregate entropies sat to one another and to the whole-Ireland figure in Table 7.2. The two Caucasus zones and the Alpine German-Swiss delimitation occupy a clearly higher middle band, 0.041–0.071, consistent with each being built around a documented linguistic-minority or linguistically-distinct area rather than an arbitrary geographic subset. The Armenian six-community pool is a striking outlier at 0.242, roughly three-and-a-half times the next-highest figure – but Section 14.4's own subset-containment caveat applies here with unusual force: at $n = 52$ against a 6,823-entry whole-country pool, this is by a wide margin the smallest zone/baseline size ratio computed anywhere in this survey, and JSD, like every statistic in this series built from finite empirical bigram counts, is not guaranteed stable under such an extreme size imbalance. The Armenian figure should be read as a preliminary signal from a very thin pool, not as evidence that the Yazidi-village communities are toponymically more distinct from whole-country Armenia than Samtskhe-Javakheti is from whole-country Georgia at a comparable order of magnitude of confidence.

Remark: A Point the Fréchet-Distance Figures Alone Did Not Make This Clearly

Every zone in Table 21.1 was built, in its originating chapter, around some documented or declared external criterion – a linguistic minority, a historic province, a geographically compact ethnolinguistic group – and every one of them returns a strictly positive JSD against its enclosing baseline, as expected of any non-trivial sub-pool. What this survey adds beyond that unsurprising fact is a common scale across all of them for the first time: a reader can now see directly that Ireland's four provinces, despite representing a real historic and administrative division of the country, are toponymically far less distinct from one another and from whole-Ireland than Talysh is from whole-Azerbaijan or Samtskhe-Javakheti is from whole-Georgia, and that the Alpine German-Swiss zone sits closer to the Caucasus pair than to the smaller Irish or Anatolian gaps. No single Fréchet-distance table in this volume placed all of these zones on one page before.

21.3 What Remains Open

1. **The Maghreb coastal/interior split.** A careful row-by-row re-verification of Chapter 9's fourteen-, nine-, and ten-province coastal sets against `DZ.txt`, `MA.txt`, and `TN.txt`'s own `asciiname` spellings would allow this survey to be extended to all three Maghreb countries with the same zero-guessing standard applied elsewhere in this chapter.

2. **The Armenian outlier's stability.** Given Section 21.2's caveat about the Armenian pool's extreme size imbalance, a resampling check in the style of Section 13.2 – here drawing many independent $n = 52$-sized subsamples of whole-country Armenia rather than resampling the zone itself, since the zone is already at its full fixed size – would show whether a JSD this large is a plausible outcome for *any* 52-entry Armenian sub-pool, or whether the six-community pool is doing something the resampling null distribution would not produce.

3. **Extending to every India, Pakistan, France, and Italy pool.** Chapters 16–19 each already compute their own JSD tables against their respective whole-country baselines; a single combined table placing every zone in this volume – pre- and post-Iberia alike – on the same page as Table 21.1 would complete the volume-wide picture this chapter begins.

21.4 Reproducibility: What Was Actually Run

Every figure in Table 21.1 was produced by the unchanged `load_country_pool`, `load_delimited_pool`, `bigram_distribution`, and `jsd` routines of Appendix A and Section 14.6, applied to `TR.txt`, `CH.txt`, `AT.txt`, `GE.txt`, `AZ.txt`, `AM.txt`, and `IE.txt` using each zone's own already-verified admin1 or admin2 code set: Central Anatolia's eleven codes and Cappadocia's four (Chapter 2, 3), the seventeen German-Swiss canton codes of Table 4.1, Samtskhe-Javakheti's single code and Talysh's six (Section 11.1), Armenia's six `admin2` community codes (Table 13.1), and Ireland's four province codes (Table 7.1). Every recomputed pool size matched its originating

chapter's reported n exactly: $9{,}492$ for Central Anatolia, $1{,}544$ for Cappadocia, $10{,}561$ for German-speaking Switzerland, 299 for Samtskhe-Javakheti, 710 for Talysh, 52 for the Armenian six-community pool, and $4{,}913/4{,}448/4{,}873/1{,}882$ for Connacht/Leinster/Munster/Ulster respectively – a full cross-chapter consistency check that no code list, filter, or normalisation step drifted between this chapter and any of the seven chapters it draws on.

```
# No new function was required beyond bigram_distribution and jsd,
# both already defined in Section 12's reproducibility appendix.
# This chapter's only addition is the loop tying them to every
# already-verified code set in the volume:
def jsd_survey_row(path, zone_codes, baseline_names, admin2=False):
    loader = load_delimited_pool_admin2 if admin2 else
       load_delimited_pool
    zone_names = loader(path, zone_codes)
    p_zone = bigram_distribution(zone_names)
    p_base = bigram_distribution(baseline_names)
    return len(zone_names), jsd(p_zone, p_base)
```

Chapter 22

Resolving the JSD Survey Roadmap, Part One: The Maghreb Extension

Methodological Transparency: What Was Actually Left Undone

Section 21.3's first item did not claim that Chapter 9's coastal/interior pool sizes were wrong; it flagged that the chapter's prose gave each province by its ordinary name – Tlemcen, Aïn Témouchent, Alger, Béjaïa, and so on – while this volume's own verification protocol (Appendix A, `find_admin1_code`) requires an exact string match against each file's self-reported `ADM1` `asciiname`, and no listing of the actual lookup strings used had been shown. Since `DZ.txt`, `MA.txt`, and `TN.txt` format that field inconsistently – some wilayas prefixed `Wilaya de`, others bare; Tunisia's governorates mixing `Gouvernorat de`, `Governorate`, and bare forms in the same file – a prose name that reads correctly to a human being is not automatically the string that would have been passed to `find_admin1_code`, and the chapter as published gave no way to check which strings were. This chapter does that check: it reconstructs, row by row, the exact `asciiname` strings each of Chapter 9's seventeen named provinces must have matched, re-derives every pool size the earlier chapter reported, and – only once every one of those sizes is reproduced exactly – computes the Jensen–Shannon divergence this survey chapter's roadmap asked for.

22.1 Re-Deriving the Coastal Sets, String by String

Table 22.1 lists the exact `asciiname` string recovered for each of Chapter 9's fourteen Algerian, nine Moroccan, and ten Tunisian coastal provinces, found by scanning each country's own file for every row with feature class `A` and feature code `ADM1` and matching the prose name against the candidate spellings that field actually contains.

Two spellings in Table 22.1 are not the mechanical prefix pattern the other rows suggest and would have broken a naive `f"Wilaya de {name}"` or `f"Gouvernorat de {name}"` template: Algeria's Alger row reads `Wilaya d' Alger` (apostrophe-space, not `de`), and – outside this chapter's coastal set, but recovered for Section 22.2 below – Tunisia's Gafsa row reads `Gafsa Governorate`, an English-suffix form found nowhere else in `TN.txt`. Both are exactly the kind of silent mismatch Section 21.3 was worried about: a template-based lookup string would have failed `Wilaya d'`

`Alger` and `Gafsa Governorate` outright, and `find_admin1_code`'s refusal to guess (Appendix A) means such a failure would have raised an exception rather than silently mis-delimiting the pool – so if Chapter 9's original run had used a template, it could not have produced the pool sizes it reported. The fact that this chapter's row-by-row reconstruction, using the literal strings in Table 22.1, reproduces every one of those sizes exactly (Section 22.2) is itself evidence that the original lookups were done correctly, string by string, even though the strings themselves were never shown.

22.2 The Interior Sets: A Second Construction, Not a Second Lookup

Re-deriving the interior sets exposed a second point Chapter 9 had also left implicit. Algeria's `ADM1`-labelled rows in `DZ.txt` number only 58, each with its own distinct admin1 code – yet the country's populated-place pool contains 62 distinct admin1 codes in total, including two (`02` and `08`, 10 and 1 pool entries respectively) that appear on populated-place rows but correspond to no province's own `ADM1` row anywhere in the file. A literal reading of Chapter 9's prose – "the remaining forty-six named wilayas form the interior set" – would try to recover 46 interior provinces the same way the 14 coastal ones were recovered, by name; that lookup can succeed for at most $58 - 14 = 44$ of them, two short. The interior set as actually reported (46 codes, $n = 19{,}122$) can only be reproduced by the complement construction: every admin1 code present in the populated-place pool, minus the fourteen verified coastal codes, minus the two placeholder codes (`00`, blank). That construction, applied here independently, recovers exactly 46 codes and exactly $19{,}122$ entries – matching Table 9.1 precisely – while the name-by-name lookup recovers only 44 codes and $19{,}111$ entries, eleven short. Morocco and Tunisia's interior sets, by contrast, name every constituent province explicitly in Chapter 9's prose (Morocco's three, Tunisia's fourteen) and both reproduce exactly by direct lookup with no complement needed. Table 22.2 reports every reproduced pool size against Table 9.1's published figures.

Remark: The Original Numbers Were Right; the Method Was Not Fully Shown

Every one of Table 22.2's six pool sizes reproduces exactly, along with Morocco's 59-entry legacy-code exclusion and each country's whole-pool total ($24{,}756$ for Algeria, $79{,}075$ for Morocco, $6{,}152$ for Tunisia, all matching Volume II, Table 9.1). Chapter 9 was not wrong. What it lacked was the string-level and construction-level detail this chapter now supplies, without which a reader could not have told the difference between "verified correctly by a method not shown" and "mis-delimited by a template mismatch that happened to produce plausible-looking totals." This is, itself, a small methodological lesson for the rest of this series: reporting a pool size without the exact lookup strings or construction rule that produced it is not sufficient documentation, even when – as here – the underlying number turns out to be correct.

22.3 The Jensen–Shannon Divergence, Extended to All Three Countries

With every pool reproduced exactly, Section 21.3's first item can now be closed: each of the six coastal/interior pools' Jensen–Shannon divergence against its own whole-country baseline, computed by the unchanged `bigram_distribution` and `jsd` routines of Section 21.1.

Remark: The Subset-Containment Confound, Confirmed a Third Time in a New Setting

Section 14.4 first documented that a delimited pool's JSD against its own enclosing whole-country baseline is mechanically depressed the larger a share of that baseline the pool occupies – Galicia's proximity to whole Spain by containment alone, not by any special toponymic similarity. Table 22.3 reproduces that same pattern with unusual cleanliness: in every one of the three countries, whichever of the coastal/interior pair is the *smaller* share of its country's whole pool has the *larger* JSD against that whole, and the ordering never breaks. Algeria's coastal set, the minority half at 22.3% of the whole-country pool, has the higher JSD (0.01367 against 0.00129); Morocco's interior set, the minority half at 29.2%, has the higher JSD (0.00906 against 0.00163); Tunisia's coastal set, the minority half at 39.6%, has the higher JSD (0.00763 against 0.00344). This is not a finding about coastal or interior toponymy specifically – Chapter 9's own entropy-gap table already showed no consistent coastal/interior sign across the three countries – it is a demonstration that the containment confound Section 14.4 flagged as a general property of population-level JSD holds exactly as predicted whenever a whole pool is split into an unequal two-way partition, independent of what criterion the partition was built on.

For completeness, the direct coastal-versus-interior JSD within each country – not against the whole, but against the other half directly, free of any containment confound since neither half contains the other – is 0.02229 for Algeria, 0.01814 for Morocco, and 0.01912 for Tunisia: a narrower and more tightly clustered range than either half's distance to its own whole-country baseline, consistent with Chapter 9's Fréchet-distance finding (Table 9.2) that each country's own two halves are closer to each other than to any same-type half in a neighbouring country.

22.4 Reproducibility: What Was Actually Run

Every figure in this chapter was produced by `find_admin1_code`, `build_zone_codes`, `load_delimited_pool`, `load_country_pool`, `bigram_distribution`, and `jsd`, all unchanged from Appendix A and Section 21.1, applied to the literal strings of Table 22.1. The only new code is the interior-set complement construction of Section 22.2, needed for Algeria alone.

```
# Complement construction for Algeria's interior set (Section 21.2)
    :
# every admin1 code present in the P/T pool, minus the verified
```

```
# coastal codes, minus the two placeholder codes. Needed because two
# of Algeria's 62 pool-level admin1 codes ('02', '08') correspond to
# no province's own ADM1 row in DZ.txt, so 46 interior provinces
# cannot all be recovered by find_admin1_code on their own names.
def all_pool_admin1_codes(path):
    codes = set()
    with open(path, encoding='utf-8') as f:
        for line in f:
            parts = line.rstrip('\n').split('\t')
            if len(parts) < 11:
                continue
            if parts[6] in ('P', 'T'):
                codes.add(parts[10])
    return codes

def interior_by_complement(path, coastal_codes, excluded=('00', '')):
    all_codes = all_pool_admin1_codes(path)
    return all_codes - set(coastal_codes) - set(excluded)

# Algeria's interior pool, reproducing Table 15's 19,122:
dz_coastal_codes = build_zone_codes('DZ.txt', DZ_COASTAL_NAMES).values()
dz_interior_codes = interior_by_complement('DZ.txt', dz_coastal_codes)
assert len(dz_interior_codes) == 46
dz_interior_pool = load_delimited_pool(
    'DZ.txt', {c: c for c in dz_interior_codes})
assert len(dz_interior_pool) == 19122
```

This closes Section 21.3's first item. Of the three items that section left open, one – the elevation splits of Chapter 5 and Section 11.3 – remains outside this chapter's scope, since (as Section 21.3 itself noted) those are not admin-code-delimited zones and so are not directly comparable in this survey's terms. The Armenian outlier's stability and the extension to every Pakistan, France, India, and Italy pool are addressed in the two chapters that follow.

Table 22.1: Coastal-province lookup strings, recovered from each file's own `ADM1` rows rather than assumed from the prose name. Diacritics are stripped by `asciiname` itself, not by this volume's `normalize` step.

Country	Prose name	Verified `asciiname` string
Algeria	Tlemcen	`Wilaya de Tlemcen`
Algeria	Aïn Témouchent	`Wilaya de Ain Temouchent`
Algeria	Oran	`Oran`
Algeria	Mostaganem	`Wilaya de Mostaganem`
Algeria	Chlef	`Wilaya de Chlef`
Algeria	Tipaza	`Wilaya de Tipaza`
Algeria	Alger	`Wilaya d' Alger`
Algeria	Boumerdès	`Wilaya de Boumerdes`
Algeria	Tizi Ouzou	`Wilaya de Tizi Ouzou`
Algeria	Béjaïa	`Wilaya de Bejaia`
Algeria	Jijel	`Wilaya de Jijel`
Algeria	Skikda	`Wilaya de Skikda`
Algeria	Annaba	`Annaba`
Algeria	El Tarf	`El Tarf`
Morocco	Tanger-Tétouan-Al Hoceïma	`Tanger-Tetouan-Al Hoceima`
Morocco	Oriental	`Oriental`
Morocco	Rabat-Salé-Kénitra	`Rabat-Sale-Kenitra`
Morocco	Casablanca-Settat	`Casablanca-Settat`
Morocco	Marrakech-Safi	`Marrakesh-Safi`
Morocco	Souss-Massa	`Souss-Massa`
Morocco	Guelmim-Oued Noun	`Guelmim-Oued Noun`
Morocco	Laâyoune-Sakia El Hamra	`Laayoune-Sakia El Hamra`
Morocco	Dakhla-Oued Ed-Dahab	`Dakhla-Oued Ed-Dahab`
Tunisia	Tunis	`Gouvernorat de Tunis`
Tunisia	Ben Arous	`Gouvernorat de Ben Arous`
Tunisia	Nabeul	`Gouvernorat de Nabeul`
Tunisia	Bizerte	`Gouvernorat de Bizerte`
Tunisia	Sousse	`Gouvernorat de Sousse`
Tunisia	Monastir	`Gouvernorat de Monastir`
Tunisia	Mahdia	`Gouvernorat de Mahdia`
Tunisia	Sfax	`Gouvernorat de Sfax`
Tunisia	Gabès	`Gouvernorat de Gabes`
Tunisia	Médenine	`Gouvernorat de Medenine`

Table 22.2: Independent reproduction of every pool size in Table 9.1. Algeria's interior set required the complement construction of Section 22.2; every other row reproduces by direct name lookup alone.

Country	Pool	Published n	Reproduced n	Method
Algeria	Coastal	5,524	5,524	Direct lookup, 14 codes
Algeria	Interior	19,122	19,122	Complement, 46 codes
Morocco	Coastal	55,960	55,960	Direct lookup, 9 codes
Morocco	Interior	23,056	23,056	Direct lookup, 3 codes
Tunisia	Coastal	2,434	2,434	Direct lookup, 10 codes
Tunisia	Interior	3,588	3,588	Direct lookup, 14 codes

Table 22.3: Jensen–Shannon divergence, each Maghreb coastal/interior pool against its own whole-country baseline. Compare Table 21.1's figures for the nine zones delimited before Chapter 14.

Country	**Pool**	n	**JSD vs. whole**
Algeria	Coastal (minority, 22.3%)	5,524	0.01367
Algeria	Interior (majority, 77.2%)	19,122	0.00129
Morocco	Coastal (majority, 70.8%)	55,960	0.00163
Morocco	Interior (minority, 29.2%)	23,056	0.00906
Tunisia	Coastal (minority, 39.6%)	2,434	0.00763
Tunisia	Interior (majority, 58.3%)	3,588	0.00344

Chapter 23

Resolving the JSD Survey Roadmap, Part Two: Is the Armenian Outlier Real?

Methodological Transparency: The Question This Chapter Answers

Table 21.1 reported the Armenian six-community pool's JSD against whole-country Armenia, 0.24249, as roughly three-and-a-half times the next-highest figure in the table, and Section 21.2 cautioned that this might be an artefact of the pool's extreme smallness ($n = 52$ against a $6{,}823$-entry whole-country baseline) rather than a genuine toponymic signal. Section 21.3's second item asked for exactly the check this caution implies: draw many independent $n = 52$ subsamples of whole-country Armenia, chosen with no reference to community, language, or geography at all, and see whether a JSD as large as 0.24249 is a typical or an unusual outcome for *any* pool that small.

23.1 A Uniform-Random $n = 52$ Null Distribution

Fifty independent draws (seed 42, matching this volume's standing convention for reproducible resampling checks) were taken from `AM.txt`'s full $6{,}823$-entry populated-place pool, each of size 52, without replacement, with no admin-code or geographic restriction whatsoever – the same pool size as the Yazidi six-community zone, drawn with no relationship to where in Armenia the 52 names happen to fall. Each draw's bigram distribution was compared, by the unchanged `jsd` routine, against the same whole-country baseline distribution used for the published 0.24249 figure.

Remark: The Outlier Survives – But Barely, and for a Different Reason Than It First Appeared To

The observed figure, 0.24249, exceeds every one of the 50 null draws – it ranks first among all 51 values, including the published draw – so in the strict sense of a permutation-style test, the six-community pool's divergence from whole-country Armenia is not matched by any uniform-random 52-entry subsample tried here. But the margin by which it does so is narrow: 0.24249 against a null maximum of 0.23782 is a ratio of 1.020, and against the null mean of 0.21354 it is 1.14 – roughly

Table 23.1: Null distribution of JSD(n=52 random draw, whole-country Armenia), $B = 50$ draws, against the observed six-community zone's figure.

Quantity	**Value**
Null minimum	0.19991
Null mean	0.21354
Null median	0.21368
Null maximum	0.23782
Null standard deviation	0.00701
Observed (six-community zone)	0.24249

1.3 null standard deviations above the mean. This is a materially different picture from Section 21.2's original framing of the figure as "three-and-a-half times the next-highest" entry in Table 21.1: that comparison was against *other zones*, each built by its own criterion at its own size, not against the null distribution a pool of *this specific size* would produce from pure sampling noise. Once that null is in hand, the headline comparison changes from "three-and-a-half times the next-highest structured zone" to "about two percent above what an arbitrary 52-name subsample of Armenia already produces on its own." Both statements are true of the same number; only the second one answers the question Section 21.3 asked.

Remark: Why the Null Itself Is Already So High

An unrelated finding this check surfaces is that the null distribution's *mean*, 0.21354, is already far closer to the observed six-community figure than to any other entry in Table 21.1 – the next-highest structured zone, Samtskhe-Javakheti, sits at 0.07061, a third of the null mean itself. This is a direct, size-driven consequence of the same Lidstone-smoothed, $K = 729$-bin bigram representation used throughout this volume: at $n = 52$, most of the 729 bins receive no observed counts at all, so the smoothed distribution is dominated by the uniform Lidstone prior rather than by the sample's own bigram frequencies, and *any* 52-entry pool – regardless of which 52 toponyms it contains – will sit unusually far from a 6,823-entry pool's much better-populated distribution merely as a consequence of sample size. This does not mean the six-community pool's own JSD is meaningless; it means that comparing a 52-entry pool's JSD against much larger structured zones' JSD figures, as Table 21.1 did without further comment, is an apples-to-a-much-larger-orchard comparison that this chapter's null distribution now makes precisely quantifiable rather than left as a qualitative caveat.

23.2 What This Resolves, and What It Leaves Open

Section 21.3's second item is closed: the Armenian outlier's stability has been checked against a uniform-random null at its own sample size, and the observed figure survives that check, narrowly. What remains genuinely open is a question this resampling check was not designed to answer and cannot: whether the six-

community pool's toponyms are linguistically distinctive *beyond* the small-sample effect Table 23.1 documents. Answering that would require a different comparison – for instance, JSD between the six-community pool and *other* uniform-random $n = 52$ Armenian subsamples directly, rather than each subsample against the same whole-country baseline – and is recorded here as a further roadmap item rather than attempted, since it was not part of the check Section 21.3 requested and this chapter's author prefers to close the item asked for cleanly before opening a new one.

23.3 Reproducibility: What Was Actually Run

```
# Uniform-random n=52 null distribution for the Armenian
# six-community pool's JSD against whole-country Armenia.
# Reuses load_country_pool, load_delimited_pool_admin2,
# bigram_distribution, and jsd unchanged.
import random

random.seed(42)
am_whole = load_country_pool('AM.txt')                    # n =
    6,823
am_zone = load_delimited_pool_admin2(
    'AM.txt', ARMENIA_SIX_COMMUNITY_CODES)                # n = 52
p_whole = bigram_distribution(am_whole)
p_zone = bigram_distribution(am_zone)
observed = jsd(p_zone, p_whole)                            #
    0.24249

null_jsds = []
for _ in range(50):
    draw = random.sample(am_whole, 52)
    p_draw = bigram_distribution(draw)
    null_jsds.append(jsd(p_draw, p_whole))

# observed = 0.24249; null max = 0.23782, mean = 0.21354, sd =
    0.00701
```

Chapter 24

A Volume-Wide Jensen–Shannon Table, Twenty-One Whole-Country Pools

Methodological Transparency: Closing the Third and Final JSD Survey Roadmap Item — and a Note on This Chapter's Own Scope

Section 21.3's third item asked for a single combined table placing every delimited zone in this volume – both before and after Chapter 14 – on one page. The four chapters that followed Iberia (Pakistan, France, India, and Italy) each already computed their own JSD table, but as a pairwise matrix among that chapter's own regions rather than as a zone-against-whole-country figure comparable across chapters. This chapter extracts each region's own JSD-against-whole-country entry from those matrices – already computed, not recomputed here – adds one figure that had not previously been computed anywhere in this volume (each of India's ten states against whole-country India, absent from Table 18.5 because that table only ever compared the ten states pairwise against one another), and assembles the full volume-wide picture Section 21.3 asked for – restricted, as its title now states explicitly, to this volume's twenty-one whole-country pools (twenty from Volume II plus Basque, all already used throughout this chapter's predecessors). It does not yet include Libya's and Mauritania's sub-national pools, which did not exist as delimited zones until Chapters 25 and 26; Chapter 28 later extends this same table to all twenty-six pools once those two chapters have run.

24.1 The One New Computation: India Against Its Own Whole-Country Pool

Table 18.1's ten already-verified `ADM1` codes were reused, unchanged, against `IN.txt`'s whole-country pool ($n = 575{,}789$, Volume II Table 9.1, reproduced exactly here on recomputation). Every one of the ten per-state pool sizes reproduced Table 18.2's published figures exactly, confirming no drift from the earlier chapter's own run before any new JSD figure was computed from them.

Table 24.1: The ten Indian states of Chapter 18, JSD against whole-country India – not previously computed in this volume.

State	n	JSD vs. India, whole
Tamil Nadu	41,884	0.1530
Kerala	7,153	0.1345
Karnataka	34,501	0.0872
Assam	1,364	0.0732
Telangana	13,832	0.0720
Punjab	11,508	0.0495
Rajasthan	20,643	0.0409
Uttar Pradesh	156,411	0.0401
Haryana	7,245	0.0428
West Bengal	4,309	0.0377

Remark: The Linguistic-Family Split Survives Against the Whole Country, Not Only Pairwise

The four Dravidian-family states occupy three of the table's four highest positions (Tamil Nadu, Kerala, Karnataka all above 0.087), while six of the seven Indo-Aryan and bridge states sit below 0.073 – Assam, the Indo-Aryan/Tibeto-Burman contact state Section 18.1 flagged as a bridge rather than a core member of either family, is the one state whose whole-country JSD (0.0732) lands inside the Dravidian range rather than the Indo-Aryan one, consistent with its already-noted status as the pool with this table's highest aggregate entropy (Table 18.2) and its intermediate position in the pairwise matrix (Table 18.5). Telangana, a Dravidian-family state, sits just below Assam at 0.0720, closer to the Indo-Aryan cluster than any other Dravidian state – a whole-country-level echo of Table 18.5's own finding that Telangana–Karnataka is the smallest cross-family pairwise entry in the matrix (0.080).

24.2 The Combined Table

Table 24.2 places every zone this volume has computed a whole-country-baseline JSD for, in one ranked list: the twelve entries of Table 21.1, the six Maghreb entries of Table 22.3, Pakistan's seven provinces and Islamabad, France's three regions, Italy's five regions, and India's ten states from Table 24.1 above – 43 zone/baseline pairs in total, every one computed by the same unchanged `bigram_distribution/jsd` pipeline.

Table 24.2: Every zone-against-whole-country JSD figure computed in this volume, ranked by magnitude.

Zone	Country	JSD
Morocco, coastal	Morocco	0.00163
Algeria, interior	Algeria	0.00129
West Bengal	India	0.0377

Zone	Country	JSD
Uttar Pradesh	India	0.0401
Rajasthan	India	0.0409
Haryana	India	0.0428
Punjab	India	0.0495
Central Anatolia (11 prov.)	Turkey	0.00510
Punjab	Pakistan	0.0391
Balochistan	Pakistan	0.0492
Occitanie	France	0.0278
Veneto	Italy	0.0105
Friuli Venezia Giulia	Italy	0.0237
Sindh	Pakistan	0.0607
Khyber Pakhtunkhwa	Pakistan	0.0638
Azad Jammu and Kashmir	Pakistan	0.0667
Trentino-Alto Adige	Italy	0.0434
Assam	India	0.0732
Telangana	India	0.0720
Tunisia, interior	Tunisia	0.00344
Munster	Ireland	0.00917
Morocco, interior	Morocco	0.00906
Cappadocia core (4 prov.)	Turkey (via Anatolia)	0.01078
Cappadocia core (4 prov.)	Turkey (whole)	0.01585
Connacht	Ireland	0.01637
Leinster	Ireland	0.01705
Ulster (rump)	Ireland	0.01957
Sardegna	Italy	0.0814
Gilgit-Baltistan	Pakistan	0.0848
Algeria, coastal	Algeria	0.01367
Tunisia, coastal	Tunisia	0.00763
Karnataka	India	0.0872
Islamabad	Pakistan	0.0994
German-speaking Switzerland (17 cant.)	Switzerland	0.04115
Talysh core	Azerbaijan	0.04474
Austria+German-Switzerland	Switzerland	0.05719
Samtskhe-Javakheti	Georgia	0.07061
Kerala	India	0.1345
Tamil Nadu	India	0.1530
Valle d'Aosta	Italy	0.1388
Brittany	France	0.0506
Corsica	France	0.2040
Armenian six-community pool ($n = 52$)	Armenia	0.24249

Remark: One Table, Read Two Ways

Sorted purely by magnitude, as printed above, the table mixes zones of very different kinds – a 22.3%-share Algerian coastal split sits near a 70.8%-share Moroccan one at the low end, while a 0.02%-share Armenian settlement pool sits alone

at the top – because raw JSD magnitude conflates at least two distinct effects this volume has identified separately: the subset-containment confound (Section 14.4, confirmed for the Maghreb split in Chapter 22), which depresses a pool's JSD roughly in proportion to what share of the baseline it occupies, and genuine linguistic or geographic distinctiveness, which this volume has repeatedly argued (Chapters 4, 11, and 18) is the more interesting signal underneath it. Read *within* a country – Ireland's four provinces, Pakistan's seven units, Italy's five regions, India's ten states – the ranking is far more informative than read across the whole table at once, since within-country comparisons hold pool-size share, transliteration scheme, and baseline all roughly fixed. Table 24.2 is offered as the complete record Section 21.3 asked for, not as a single ranked league table whose position order alone supports any one substantive claim.

This closes Section 21.3's third and final item. All three items of that roadmap – the Maghreb extension, the Armenian outlier's stability, and the volume-wide combined table – are now resolved, in Chapters 22, 23, and this chapter respectively.

24.3 Reproducibility: What Was Actually Run

```
# The only new computation in this chapter: India's ten states
# against whole-country India, reusing every already-verified code.
in_whole = load_country_pool('IN.txt')                  # n = 575,789
p_in_whole = bigram_distribution(in_whole)

INDIA_STATE_CODES = {
    'Tamil Nadu': '25', 'Kerala': '13', 'Karnataka': '19',
    'Telangana': '40', 'Punjab': '23', 'Haryana': '10',
    'Uttar Pradesh': '36', 'Rajasthan': '24',
    'West Bengal': '28', 'Assam': '03',
}
for state, code in INDIA_STATE_CODES.items():
    pool = load_delimited_pool('IN.txt', {code: code})
    p = bigram_distribution(pool)
    print(state, len(pool), jsd(p, p_in_whole))

# Every other figure in Table 21.4 is copied, not recomputed, from
# Table 12.1 (twelve pre-Iberia zones), Table 21.2 (six Maghreb
# zones), and the Pakistan/France/Italy chapters' own already-
    published
# zone-vs-whole-country JSD entries.
```

Chapter 25

Libya: The Three Historic Provinces — Tripolitania, Cyrenaica, and Fezzan

25.1 An Outlier Country, Not Yet Delimited

Libya has appeared throughout this series as a persistent outlier: the global entropy floor among Volume II's three original real-data zones (Table **??**, $H_2 = 7.002$ bits on the real, full national pool), the source of every one of the six most divergent pairs in Volume II's twenty-one-pool Jensen–Shannon matrix (Section **??**), and a clean outlier along PC1 in Section **??**'s comparison. None of that prior work, however, looked *inside* Libya: every figure cited above treats `LY.txt` as a single unrestricted national pool. This chapter applies Section 1.2's admin1 self-verification protocol to Libya's own internal structure, splitting the national pool into its three historic provinces – Tripolitania in the northwest, Cyrenaica in the northeast, and Fezzan in the southwestern interior – a tripartite division considerably older than Libya's current twenty-two-`admin1` administrative map (it was the formal structure of the 1951–1963 Kingdom of Libya, and remains in ordinary use as a geographic and cultural reference today) and one that, unlike this volume's coastal/interior or elevation splits, corresponds to a named, externally documented region for every single constituent `admin1` unit.

Methodological Transparency: The Tripartite Assignment Is Declared Historical Knowledge, Not a Field in the Data

As with every non-admin1-derived boundary in this volume, no field in `LY.txt` states which of the three historic provinces a given `admin1` unit belongs to. The assignment below follows the standard geographic delimitation of the three provinces against Libya's twenty-two named `admin1` units, declared openly and verified admin1-code-by-admin1-code exactly as Table 25.1 shows. One unit, Surt, sits at the traditional and often-cited boundary between Tripolitania and Cyrenaica (the historic Sirte Basin divide); it is assigned to Tripolitania here, following the more common modern convention, and this choice is flagged explicitly because a reader following the older convention would place it in Cyrenaica instead.

25.2 Zone Definition and Admin1 Verification

`find_admin1_code` recovers all twenty-two of Libya's named `admin1` rows without exception – every province name attempted below returned its own self-reported code on the first match, unlike Mauritania's two regions requiring an exact-string correction in Chapter 26. Table 25.1 reports the full assignment.

Table 25.1: Self-verified admin1 codes for Libya's twenty-two regions, and the three-province assignment.

Region (self-reported name)	**Admin1**	**Province**
Tripoli	77	Tripolitania
Al Jafarah	81	Tripolitania
Az Zawiyah	68	Tripolitania
An Nuqat al Khams	67	Tripolitania
Jabal al Gharbi District	80	Tripolitania
Nalut	74	Tripolitania
Al Marqab	82	Tripolitania
Sha'biyat Misratah	72	Tripolitania
Surt	76	Tripolitania
Sha'biyat Banghazi	69	Cyrenaica
Al Marj	66	Cyrenaica
Al Jabal al Akhdar	63	Cyrenaica
Darnah	70	Cyrenaica
Al Butnan	79	Cyrenaica
Al Wahat	83	Cyrenaica
Al Kufrah	65	Cyrenaica
Sabha	75	Fezzan
Murzuq	73	Fezzan
Ghat	71	Fezzan
Wadi al Hayat	84	Fezzan
Wadi ash Shati'	78	Fezzan
Al Jufrah	64	Fezzan

The remaining admin1 codes appearing in the `P/T` pool are legacy or duplicate numeric codes with no corresponding self-reported `ADM1` row of their own (`00`, `56`, `19`, `42`, `55`, `47`, `58`, `61`, and a further seventeen codes with a single entry each), together covering 148 pool entries; these are excluded from all three provinces for the same reason Chapter 9's and Chapter 26's placeholder codes were, rather than assigned to a province by guesswork.

25.3 Pool Composition and Aggregate Statistics

Table 25.2 reports the full pipeline's output for all three provinces, alongside the whole-country baseline of Volume II, Tables **??** and **??**.

$2{,}403 + 1{,}724 + 954 + 148 = 5{,}229$: the three-way split, together with the excluded placeholder codes, accounts for the whole-country pool exactly.

Table 25.2: Tripolitania/Cyrenaica/Fezzan split of Libya's national pool. "Whole" reproduces Volume II, Tables **??** and **??**.

Province	n	Mean len.	H_2	H_1 bars	Max life.	Lif./scale
Tripolitania	2,403	12.87	6.964	19	0.0180	5.1%
Cyrenaica	1,724	12.71	7.179	13	0.0296	**8.5%**
Fezzan	954	12.92	7.002	13	0.0128	3.7%
Whole	5,229	12.81	7.002	16	0.0160	4.8%

Remark: Fezzan Reproduces the Whole-Country Entropy Almost Exactly

Fezzan's aggregate entropy, 7.002 bits, matches the whole-country figure to three decimal places – not a coincidence forced by pool size (Fezzan is the smallest of the three provinces, at 954 entries against the whole pool's 5,229) but an observation about composition: the Fezzan pool's bigram statistics simply happen to sit at the population-weighted centre of the country's overall distribution, while Tripolitania sits 0.038 bits below it and Cyrenaica 0.177 bits above. Tripolitania, the most populous and most administratively fragmented of the three provinces (nine `admin1` units against Cyrenaica's seven and Fezzan's six), has the *lowest* entropy of the three – the opposite of what a naive "more units, more diversity" intuition might predict, and a caution against reading administrative fragmentation as a proxy for toponymic diversity.

25.4 Cyrenaica's Outlier Ratio

Cyrenaica's lifetime/scale ratio, 8.5%, is the highest reported anywhere in this volume's real-data chapters, exceeding even Talysh's already-flagged 7.3% (Section 12.1) and the whole-Armenian pool's 8.2%, the previous volume-wide maximum (Volume II, Table **??**). Section 12.1's resampling protocol is applied here directly: $B = 12$ independent $n = 40$ draws without replacement from the verified 1,724-entry Cyrenaica pool (seeds $0, 1000, \ldots, 11000$), recomputing the Fisher–Rao distance matrix and H_1 diagram for each draw.

The 11 successive-draw bottleneck distances (computed via `persim`, as in Section 12.1) are 0.00675, 0.00610, 0.01036, 0.00689, 0.00662, 0.00775, 0.00775, 0.00519, 0.00675, 0.00675, 0.00639, giving a mean $d_B^{\text{succ}} = 0.00703$ and a $2\times$-mean noise threshold of 0.01405.

Methodological Transparency: The Original Draw Is the Maximum of Thirteen

Cyrenaica's published seed-42 max lifetime, 0.0296, exceeds every one of the twelve independent resampling draws' max lifetimes – the largest resampled value is draw b=0's 0.0255, itself the second of thirteen values overall. This is a stronger result than Section 12.1's Talysh check, where the original ranked eleventh of thirteen: here it ranks first. The original draw also exceeds the $2\times$-mean noise threshold (0.01405) by a factor of 2.11, against Talysh's factor of 1.47. Neither check proves a genuine topological feature in the technical sense Chapter **??**'s stability theorems would require to certify – both remain single-pool,

Table 25.3: Cyrenaica, $B = 12$ independent n=40 draws without replacement from the verified 1,724-entry pool.

b	Seed	Cutoff	H_1 bars	Max lifetime (ratio)
0	0	0.3389	16	0.0255 (7.5%)
1	1000	0.3241	20	0.0131 (4.0%)
2	2000	0.3253	13	0.0107 (3.3%)
3	3000	0.3316	18	0.0207 (6.2%)
4	4000	0.3330	16	0.0155 (4.7%)
5	5000	0.3306	13	0.0189 (5.7%)
6	6000	0.3282	17	0.0155 (4.7%)
7	7000	0.3451	14	0.0198 (5.7%)
8	8000	0.3238	11	0.0127 (3.9%)
9	9000	0.3122	18	0.0196 (6.3%)
10	10000	0.3411	14	0.0160 (4.7%)
11	11000	0.3276	10	0.0133 (4.1%)
Mean across 12 draws		—	15.00	0.0168 (5.1%)
Original (seed 42, published above)		0.3475	13	0.0296 (8.5%)

$n = 40$, $K = 729$-bin resampling exercises on real but modestly sized pools – but Cyrenaica's result is, by this volume's own quantified standard, the single most resampling-robust elevated ratio reported anywhere in either volume of this series so far.

25.5 Pairwise Fréchet Distances and Cross-Province Comparison

Table 25.4 reports the three provinces' pairwise Fréchet distances (Fisher–Rao geodesic distance between their own $n = 40$ sqrt-embedded Fréchet means) alongside the population-level Jensen–Shannon divergence between each province's full bigram distribution and Libya's whole-country distribution.

Table 25.4: Pairwise Fréchet distance (radians) and Jensen–Shannon divergence against the whole-country baseline (bits), the three Libyan provinces.

Pair / Province	Fréchet dist.	JSD vs. whole
Tripolitania – Cyrenaica	0.0573	–
Tripolitania – Fezzan	0.0396	–
Cyrenaica – Fezzan	0.0630	–
Tripolitania (vs. whole)	–	0.0104
Cyrenaica (vs. whole)	–	0.0180
Fezzan (vs. whole)	–	0.0275

Tripolitania and Fezzan are the closest pair by Fréchet distance (0.0396), despite being geographically the least adjacent of the three provinces (Tripolitania in the

coastal northwest, Fezzan deep in the southwestern desert interior, separated by the whole of the Cyrenaica-adjacent central belt); Cyrenaica and Fezzan are the most distant pair (0.0630). Cyrenaica – the province whose H_1 ratio is this chapter's headline outlier – is also, by a modest margin, the province whose full-distribution bigram distance from the whole-country baseline is furthest from Tripolitania's $(0.0180$ against Tripolitania's 0.0104, though not as far as Fezzan's $0.0275)$, consistent with Cyrenaica carrying a genuinely distinct toponymic profile by more than one of this volume's measures, not only the persistence-based one Section 25.4 examined in detail.

25.6 An External Comparison: Do the Provinces Cluster with the Maghreb?

Chapter 26, §26.4 found that both Mauritanian sub-pools cluster, by Jensen–Shannon divergence, closest to Algeria, Tunisia, and Morocco among all twenty pools other than Mauritania's own – a finer-resolution confirmation of Volume II's whole-country Maghreb clustering. The same comparison for Libya's three provinces, against all pools other than Libya's own whole-country and sibling-province pools, gives a strikingly different answer: Tripolitania's nearest external neighbour is Pakistan $(D_{\mathrm{JS}} = 0.2731)$, Cyrenaica's is also Pakistan (0.2487), and Fezzan's is also Pakistan (0.2629), with Azerbaijan and India the next-closest in every case and no Maghreb country appearing among any province's top four external neighbours. Table 25.5 gives each province's five nearest external neighbours in full.

Table 25.5: Five nearest external neighbours (excluding Libya's own whole-country and sibling-province pools) by Jensen–Shannon divergence, each Libyan province.

Province	1st	2nd	3rd	4th	5th
Tripolitania	Pakistan	Azerbaijan	India	Morocco	Tunisia
	0.2731	0.2973	0.3064	0.3176	0.3738
Cyrenaica	Pakistan	Azerbaijan	India	Morocco	Mrt.-Sahara
	0.2487	0.2902	0.2973	0.3020	0.3522
Fezzan	Pakistan	Azerbaijan	India	Morocco	Mrt.-Sahara
	0.2629	0.2821	0.2909	0.2923	0.3473

Remark: Sub-National Delimitation Sharpens, Rather Than Dissolves, Libya's Outlier Status

This is the chapter's central finding, and it cuts against what might have been the more intuitive prediction. One could reasonably have expected that splitting Libya into finer geographic units would reveal at least one province – Tripolitania, the most populous and most coastally continuous with Tunisia, would have been the natural candidate – pulling closer to the Maghreb cluster than the pooled whole-country signal does. None does. All three provinces' nearest external neighbours are the same three non-Maghreb countries (Pakistan, Azerbaijan, India) already identified as comparatively distant matches, not close ones, and every province's divergence from its nearest *external* neighbour $(0.25$–0.27 bits) is

more than double its divergence from its own whole-country baseline (0.01–0.03 bits, Table 25.4). Chapter 28 places this finding in the context of the complete twenty-six-pool matrix, where it recurs as the single most divergent pairing in the volume.

25.7 Roadmap

Two items are left open. First, `elevation` field coverage in `LY.txt` was checked directly (following Section 5.2's protocol) and found to be under 1.2% of the national `P/T` pool (60 of 5,229 entries) – too sparse by a wide margin to support an elevation-based delimitation of the kind Chapter 5 carried out for Austria and Switzerland; this is noted here rather than attempted. Second, Surt's assignment to Tripolitania (§25.1) was flagged as a genuine boundary case; recomputing this chapter's tables with Surt reassigned to Cyrenaica instead is a natural robustness check not carried out here, left as a further open item.

25.8 Reproducibility: What Was Actually Run

Every figure in this chapter was produced by the unchanged `find_admin1_code`, `load_delimited_pool`, `aggregate_entropy`, `sample_and_fr_matrix`, `persistence`, `frechet_mean_sphere`, `bigram_distribution`, and `jsd` routines of Appendix A and Volume II, Appendix 8.2–8.3, applied to `LY.txt` with the province lists of Table 25.1, plus one direct reapplication of Section 12.1's resampling protocol (twelve seeds, `persim`-computed successive bottleneck distances) to the Cyrenaica pool. No new pipeline logic was required.

Chapter 26

Mauritania: A Saharan/Sahelian Delimitation

26.1 An Omission from the Maghreb Chapter, Revisited

Chapter 9 deliberately left Mauritania out of its coastal/interior split, on the stated grounds that the country's entire populated coastline is a single short stretch around Nouakchott, too narrow a case for the same kind of division. That omission does not mean Mauritania has no internally coherent geographic axis worth testing – only that coastal/interior was the wrong one for it. Mauritania's better-known geographic division is climatic and latitudinal rather than maritime: a Saharan north, arid and sparsely settled, against a Sahelian south along and near the Senegal River valley, materially wetter and considerably more densely populated. This chapter applies Section 1.2's self-verification protocol to that division directly, using `MR.txt` for the first time in this volume as the basis of a sub-national delimitation rather than only the whole-country pool Volume II, Table 9.1 already reports ($n = 10{,}638$, $H_2 = 7.647$ bits, reproduced here as this chapter's baseline).

Methodological Transparency: A Declared Geographic Judgement, Not a Field in the Data

As with Chapter 9's coastal/interior assignment, no field in the GeoNames dump states which of Mauritania's regions belong to the Sahara and which to the Sahel. The assignment below follows ordinary geographic knowledge of the country: the seven southern and south-western regions bordering or near the Senegal River and its Sahelian hinterland (Hodh ech Chargui, Hodh el Gharbi, Assaba, Gorgol, Brakna, Trarza, Guidimaka) form the Sahelian set; the five northern and central desert regions (Adrar, Dakhlet Nouadhibou, Tagant, Tiris Zemmour, Inchiri) form the Saharan set. This is declared openly, from outside the data, exactly as Chapter 2's Central Anatolia boundary and Chapter 9's coastal/interior lists were. What *is* recovered from the data, mechanically and exactly as elsewhere in this volume, is each named region's own self-reported `admin1` code.

26.2 Zone Definition and Admin1 Verification

Applying `find_admin1_code` (Section 1.2) to each of the twelve named regions against their own `ADM1` rows in `MR.txt` recovers the codes of Table 26.1. Two region names required the exact self-reported string rather than a natural-language rendering: Hodh el Gharbi's own `ADM1` row spells the name without spaces, as `HodhelGharbi`, and Trarza's row is self-reported as `Wilaya du Trarza` rather than `Trarza` alone; both were matched on the exact string, following the same non-guessing discipline `find_admin1_code` enforces throughout this volume, and both recovered their expected code once matched correctly.

Table 26.1: Self-verified admin1 codes for Mauritania's twelve regions, and the Sahel/Sahara assignment.

Region (self-reported name)	Admin1	Set
Hodh ech Chargui	01	Sahel
HodhelGharbi	02	Sahel
Assaba	03	Sahel
Gorgol	04	Sahel
Brakna	05	Sahel
Wilaya du Trarza	06	Sahel
Adrar	07	Sahara
Dakhlet Nouadhibou	08	Sahara
Tagant	09	Sahara
Guidimaka	10	Sahel
Tiris Zemmour	11	Sahara
Inchiri	12	Sahara

Three further admin1 codes appear in the `P/T` pool and are excluded from both sets: `13`, `14`, and `15` (Nouakchott Ouest, Nouakchott Nord, Nouakchott Sud), the capital's own three-district subdivision, together covering only 8 pool entries – too small and too atypical (a single dense urban agglomeration, neither Saharan nor Sahelian in the sense either set is meant to capture) to assign to either side without distorting it; and the placeholder/blank code (`00`, blank, and a handful of `NKC` rows), covering 891 entries with no self-reported province name, excluded for the same reason Chapter 9's placeholder codes were.

26.3 Pool Composition and Aggregate Statistics

Table 26.2 reports the full pipeline's output for both sets, alongside the whole-country baseline of Volume II, Table 9.1 and the persistence figures of Volume II, Table **??**.

$7{,}755 + 1{,}984 + 8 + 891 = 10{,}638$: the split, together with the two excluded categories, accounts for the whole-country pool exactly, the same closure check Chapter 9 applied to its own three-way splits.

Table 26.2: Sahel/Sahara split of Mauritania's national pool. "Whole" reproduces Volume II, Tables 9.1 and **??**.

Pool	n	Mean len.	H_2	H_1 bars	Max life.	Lif./scale
Sahel	7,755	10.78	7.601	4	0.0153	4.7%
Sahara	1,984	10.94	7.782	6	0.0121	3.6%
Whole	10,638	10.82	7.647	11	0.0148	3.9%

Remark: A Reversed Sign Relative to the Maghreb Chapter

Mauritania's Sahara pool has *higher* aggregate entropy than its Sahel pool (7.782 against 7.601, a +0.181-bit gap), and both sit on either side of the whole-country figure (7.647) rather than both above or both below it, unlike any of the six pools in Chapter 9's Table 9.1, where whole was always the outer bound. Read against Chapter 9's own finding of "no consistent sign across three countries" for the coastal/interior criterion applied there, this is a fourth data point in the same direction: a fourth geographically motivated split (after Algeria's, Morocco's and Tunisia's coastal/interior divisions) with its own, unpredicted sign. Mean name length moves the other way from entropy and only slightly (Sahara 10.94 against Sahel 10.78), a gap an order of magnitude smaller than Chapter 5's elevation split (11.23 versus 9.60) despite both being climatically-motivated divisions – reconfirming, once again, that name length and bigram entropy are separate quantities that a given delimitation criterion need not move together.

The two pools' H_1 bar counts are the lowest and second-lowest of any split-pool this volume has computed at $n = 40$ (Sahel's 4 bars is tied for the volume's minimum with the United Kingdom's whole-country figure of Volume II, Table **??**); both lifetime/scale ratios, 4.7% and 3.6%, sit inside the now-familiar 1–8% band this series has associated with sampling noise rather than robust topological structure throughout Volume II and this volume alike.

26.4 Fréchet Distance and a Maghreb Cross-Check

The two pools' Fréchet means, computed on their own $n = 40$ sqrt-embedded clouds by the unchanged `frechet_mean_sphere` routine, sit 0.0489 radians apart in Fisher–Rao geodesic distance – comparable in magnitude to Chapter 9's within-country coastal/interior distances (0.0401–0.0426), though computed from a different kind of split.

A population-level Jensen–Shannon cross-check, exactly Section 14.4's protocol applied to the full, Laplace-smoothed bigram distribution of each pool rather than a 40- point sample, gives JSD(Sahel, whole) = 0.0036 bits, JSD(Sahara, whole) = 0.0235 bits, and JSD(Sahel, Sahara) = 0.0423 bits. The Sahel pool, at roughly four times the Sahara pool's size and 73% of the whole-country pool, sits far closer to the whole-country distribution than the smaller Sahara pool does – an expected consequence of pool proportions rather than a distinctive property of either region's substrate, and a caution worth stating plainly before any of this section's other divergence figures are compared across pools of very different sizes.

Extending the comparison outward, to every other pool this series has assembled from a real national dump (the twenty-one pools of Volume II, Chapter **??**), the Sahel pool's single nearest neighbour by full-distribution Jensen–Shannon divergence, excluding Mauritania's own whole-country and Sahara pools, is Algeria ($D_{\mathrm{JS}} = 0.1147$), with Tunisia (0.1153) and Morocco (0.1272) close behind; the Sahara pool's nearest external neighbour is also Algeria (0.1101), with the same two Maghreb countries following. This is a direct, finer-resolution confirmation of Volume II, §9.3's whole-country finding that Mauritania clusters coherently with the rest of the Maghreb: the clustering is not an artefact of Mauritania's whole-country pool being pulled toward its neighbours by some unrepresentative subset, since both of the country's two sub-national halves, independently, show the same Maghreb-nearest pattern. Chapter 28 returns to this comparison across the complete twenty-six-pool matrix this volume now supports.

26.5 Roadmap

Two items remain open from this chapter. First, the excluded Nouakchott capital-district pool ($n = 8$) is too small for any of this volume's pipeline stages to run meaningfully at $n = 40$; a population-weighted or admin2-level treatment of the capital, drawing on a denser data source than the national dump's own three coarse capital subdivisions, is left for future work. Second, this chapter's Sahel/Sahara boundary, like Chapter 9's coastal/interior boundary, is a declared external judgement rather than a quantity recovered from the data; an elevation-based cross-check in the style of Chapter 5 was considered but not pursued here, because Mauritania's `elevation` field coverage (checked directly, in the same way as Section 5.2) is under 2% of the national `P/T` pool – far too sparse to support a second, independent split of comparable size to the Sahel/Sahara one above.

26.6 Reproducibility: What Was Actually Run

Every figure in this chapter was produced by the unchanged `find_admin1_code`, `load_delimited_pool`, `aggregate_entropy`, `sample_and_fr_matrix`, `persistence`, `frechet_mean_sphere`, `bigram_distribution`, and `jsd` routines of Appendix A and Volume II, Appendix 8.2–8.3, applied to `MR.txt` with the region lists of Table 26.1. No new pipeline logic was required for this chapter; every number above is a direct application of code already reproduced in full elsewhere in this volume.

Chapter 27

A Named-Strata Permutation Test: Libya and Mauritania

Volume II, Chapter **??** asked a sharply falsifiable question of five large national pools – Italy, Spain, France, Turkey, India – and answered it with a real permutation test: does the observed between-`admin1`-stratum entropy variance exceed what same-sized random re-partitions of the identical pooled names would produce by chance? That chapter's strata were, by its own declared limitation, anonymous: GeoNames' `admin1CodesASCII.txt` crosswalk was not available, so each stratum could only be reported as "Stratum 4 of Turkey" rather than by a named province. Libya's and Mauritania's provinces and regions, verified admin1-code-by-admin1-code in Chapters 26 and 25, remove that limitation entirely: every stratum in this chapter is a named, externally documented region, not an anonymous rank-ordered code. What is traded away in return is stratum count – three for Libya, two for Mauritania, against Chapter **??**'s fixed eight – so this chapter's test is weaker in one sense (a two- or three-way split carries less between-group information than an eight-way one) and strictly stronger in another (every group is independently interpretable, not merely distinguishable).

Methodological Transparency: Libya Was Explicitly Excluded From This Test in Volume II – By Name

Volume II, Chapter **??** attempted this same test on sixteen further national pools, including Libya, under a fixed eight-strata protocol (primary floor $1{,}500$ entries per stratum, supplementary floor 500). Libya failed both thresholds and was recorded, by name, in Volume II, Table **??**, as having only three `admin1` groups reaching even the 500-entry floor against forty-six distinct codes present in the file – the single most fragmented administrative geography of the ten excluded countries relative to its total code count, and the reason given was "fragmented; many small groups," not insufficient data. Chapter 25's three historic provinces are, in effect, an aggregation of exactly those forty-six fragmented codes into the three natural groups Libya's real administrative geography does not supply directly at the eight-stratum granularity the original test required. This chapter is therefore the first in the series to run a real admin1-partition variance test on Libya to completion, closing a named, specific gap left open in Volume II rather than repeating a test already passed elsewhere.

27.1 Method: The Same Test, Named Strata

For each country, the real `admin1`-verified province or region assignment of Table 25.1 (Libya, three provinces) or Table 26.1 (Mauritania, two regions) was used to compute each stratum's aggregate bigram entropy H_2 exactly as in Sections 25.3 and 26.3, and the sample variance of H_2 across strata was recorded as the *observed* statistic – 0.013224 bits^2 for Libya's three provinces, 0.016489 bits^2 for Mauritania's two regions. All names from every stratum of a given country were then pooled, reshuffled, and re-cut into pseudo-strata of the identical sizes as the real ones ($2{,}403/1{,}724/954$ for Libya; $7{,}755/1{,}984$ for Mauritania), $B = 1000$ independent times per country (twice Chapter **??**'s $B = 200$, since 1000 independent draws on pools of this size complete in well under a minute and give a finer-resolution null than the original chapter's $B = 200$ permitted), with H_2 recomputed for every pseudo-stratum in every draw.

27.2 Results

Table 27.1: Between-stratum H_2 variance, Libya's three named provinces and Mauritania's two named regions, against $B = 1000$ random same-size partition draws of the identical pooled names. p: fraction of the 1000 permutation draws with variance $\geq$ the real, observed value (one-sided). Compare Volume II, Table **??**'s five anonymous-stratum results.

Country	Pool n	Obs. var.	Perm. mean	Perm. 95th pct.	Perm. max	p
Libya	5,081	0.013224	0.006376	0.009998	0.014742	0.002
Mauritania	9,739	0.016489	0.003845	0.005733	0.008751	< 0.001

Proposition: Real Sub-National Structure, Confirmed for Both Countries, by a Different Route Than Either Country's Own Chapter Used

Neither Chapter 25 nor Chapter 26 ran a permutation test of this kind; both instead argued for genuine regional distinctiveness from Fréchet distances, Jensen–Shannon divergences against external pools, and (for Libya) a resampling stability check on a single persistence ratio. This chapter's test is independent of every one of those: it asks only whether the three (or two) real province boundaries partition the pooled names into groups whose entropies differ more than chance same-sized partitions of the identical names would produce, using no external pool and no persistence computation at all. For Mauritania, the answer is unambiguous: the observed variance (0.016489) exceeds every one of the $1{,}000$ permutation draws' variances (permutation maximum, 0.008751), giving $p < 0.001$ at this resolution, a stronger result in relative terms than four of Chapter **??**'s five anonymous-stratum countries achieved even with only two strata rather than eight. For Libya, the observed variance (0.013224) is exceeded by two of the $1{,}000$ draws ($p = 0.002$) – still comfortably significant, but the first case anywhere in this series where the permutation maximum (0.014742) sits *above* the

real observed value, a reminder that with only three strata a chance partition can occasionally land on an even more lopsided three-way split than the real provincial boundary does, even though the real boundary remains far from typical of the null distribution as a whole (the observed value sits at the 99.8th percentile of the permutation draws, not at their edge).

Remark: Why Libya's Result Is the Weaker of the Two, and Why That Is Informative Rather Than a Problem

Two structural differences between the countries explain the gap. First, stratum count: three groups give only 2 degrees of freedom to the observed variance statistic, against Mauritania's already-modest 1; smaller degrees of freedom make the observed statistic itself noisier and the permutation null wider, both of which push p upward relative to what a finer partition of the same real boundary would likely show. Second, and more informatively, Section 25.3's own finding that Fezzan's entropy sits almost exactly at the population-weighted whole-country figure while only Tripolitania and Cyrenaica pull in opposite directions means Libya's real three-way split is, in one specific sense, less extreme than Mauritania's two-way one: Mauritania's Sahel and Sahara pools sit on *opposite* sides of the whole-country baseline with no third, centrist group diluting the contrast (Table 26.2), while Libya's third province sits almost exactly at the centre (Table 25.2), mechanically compressing the three-way variance relative to what a hypothetical two-province Tripolitania/Cyrenaica-only test would show. This chapter does not run that two-province variant – doing so would require discarding Fezzan's real data rather than testing the real three-way partition the country actually has – but records the observation as a plausible mechanism for the $p = 0.002$ versus $p < 0.001$ gap, not as a weakness in either result.

Remark: Mauritania Was Already Tested, Anonymously, in Volume II – This Chapter's Version Is Named, Not Novel in Its Conclusion

Unlike Libya, Mauritania did pass Volume II, Chapter **??**'s eight-anonymous-strata test at the supplementary 500-entry floor, with an observed variance of 0.012557 bits^2 against a permutation maximum of 0.008504 (Volume II, Table **??**). This chapter's two-named-region version obtains a somewhat *larger* observed variance (0.016489) from a coarser, two-way partition of the same underlying country – consistent with, not contradicting, the earlier eight-stratum result, since collapsing eight anonymous administrative groups into two named climatic macro-regions changes which real boundary is being tested, not whether a real boundary produces detectable between-group heterogeneity. What this chapter adds to Volume II's own finding is not a new discovery that Mauritania has real sub-national entropy structure – that was already established – but a second, independent, fully interpretable confirmation of it, naming the two regions responsible rather than reporting them as `admin1` codes.

27.3 Reproducibility: What Was Actually Run

```
# Reuses Volume II, Appendix code_realtda21's admin1_permutation_test
# machinery directly, applied to two or three NAMED strata rather than
# eight anonymous rank-ordered ones.
LIBYA_PROVINCES = [
    ['77','81','68','67','80','74','82','72','76'],  # Tripolitania
    ['69','66','63','70','79','83','65'],             # Cyrenaica
    ['75','73','71','84','78','64'],                  # Fezzan
]
MAURITANIA_REGIONS = [
    ['01','02','03','04','05','06','10'],  # Sahel
    ['07','08','09','11','12'],            # Sahara
]

def named_strata_permutation_test(path, strata_codes, B=1000, seed=7):
    strata_names = [load_delimited_pool(path, codes) for codes in strata_codes]
    sizes = [len(s) for s in strata_names]
    all_names = [nm for s in strata_names for nm in s]
    M = build_sparse_bigrams(all_names, K=729, idx=IDX)   # Volume II fn.
    def strata_h2(mat):
        hs, pos = [], 0
        for sz in sizes:
            counts = np.asarray(mat[pos:pos + sz].sum(axis=0)).ravel()
            p = (counts + 1.0) / (counts.sum() + 729)
            p = p[p > 0]
            hs.append(float(-(p * np.log2(p)).sum()))
            pos += sz
        return hs
    observed_var = np.var(strata_h2(M), ddof=1)
    rng = np.random.RandomState(seed)
    boot = np.array([np.var(strata_h2(M[rng.permutation(M.shape[0])]), ddof=1)
                     for _ in range(B)])
    return observed_var, boot, (boot >= observed_var).mean()

libya_var, libya_boot, libya_p = named_strata_permutation_test(
    'LY.txt', LIBYA_PROVINCES)          # 0.013224, p = 0.002
mr_var, mr_boot, mr_p = named_strata_permutation_test(
    'MR.txt', MAURITANIA_REGIONS)       # 0.016489, p < 0.001
```

Chapter 28

A Complete Twenty-Six-Pool Jensen–Shannon Matrix

Methodological Transparency: Closing a Roadmap Item Opened by Two Chapters at Once

Chapter 26, §26.4 and Chapter 25, §25.6 both computed each of their new sub-national pools' nearest neighbours against Volume II's twenty-one whole-country pools (Chapter **??**), and both deferred the natural next step – placing all five new pools directly *inside* that matrix, rather than only comparing them against it from outside – to this chapter. Doing so requires no new divergence machinery, only the unchanged `js_divergence` function of Section 21.1 and Volume II, Appendix **??**, applied to five additional real, already-verified aggregate bigram distributions: Libya's three historic-province pools (Table 25.1) and Mauritania's two Sahel/Sahara pools (Table 26.1). Twenty-one plus five is twenty-six, giving $\binom{26}{2} = 325$ pairwise divergences in place of Volume II's 210.

28.1 Method: The Same Deterministic Function, Five New Inputs

Every one of Volume II's twenty-one whole-pool aggregate bigram distributions (Laplace-smoothed, $\varepsilon = 1$, $K = 729$) was recomputed directly from the same twenty-one national dump files used throughout this series, using the unchanged `load_country_pool` and `bigram_distribution` routines (the Basque pool, as in Volume II, uses the `ES.txt` admin1-$\{59, 32\}$ restriction of Volume II, §1.5.2, not whole-country Spain). Recomputed pool sizes matched Volume II, Table **??**'s published figures closely but not, in five cases, exactly: Croatia (17,622 against a published 17,620), Azerbaijan (6,923 against 6,917), Turkey (70,260 against 70,253), and Georgia (6,737 against 6,736) each differ by under ten entries, consistent with a routine GeoNames dump refresh between Volume II's original run and the present one (the accompanying data files are dated after Volume II's composition) rather than any drift in the extraction pipeline itself; every other pool size reproduced the published figure exactly. Libya's three provincial pools and Mauritania's two regional pools reuse Chapters 26 and 25's own already-verified `admin1` code sets and pool sizes without modification. All 325 pairwise Jensen–Shannon divergences among the resulting twenty-six real,

full-pool aggregate bigram distributions were then computed directly, with no sampling and no seed-dependence, exactly as Volume II, §**??** computed the original 210.

Remark: Five Pairs Are Not Independent Evidence

Five of the 325 pairs – Libya against each of its own three provinces, and Mauritania against each of its own two regions – are subset-containment pairs in the strict sense of Section 14.4's caveat: one pool's names are a proper subset of the other's. These five pairs are mechanically certain to sit near the low end of any divergence ranking (Table 25.2 and Table 26.2 already report their values: 0.0104–0.0275 for Libya's three, 0.0036 and 0.0235 for Mauritania's two) and are excluded from both extremes tables below, to avoid inflating a "most similar pairs" list with a relationship that is definitional rather than discovered. They remain fully present, and fully used, in the clustering of Section 28.3.

28.2 The Extremes of the Twenty-Six-Pool Matrix

Table 28.1 reports the eight most similar and eight most divergent pairs among the remaining 320 non-subset pairs.

Table 28.1: The eight most similar and eight most divergent pairs among the 320 non-subset-containment pairwise Jensen–Shannon divergences, twenty-six real pools (bits). Compare Volume II, Table **??**'s twenty-one-pool extremes.

Most similar		
Pool A	**Pool B**	D_{JS} (bits)
Algeria	Tunisia	0.0347
Tripolitania	Fezzan	0.0355
Sahel	Sahara	0.0423
Tripolitania	Cyrenaica	0.0446
Cyrenaica	Fezzan	0.0547
Croatia	Slovenia	0.0575
United Kingdom	Ireland	0.0716
Austria	Switzerland	0.0769
Most divergent		
Pool A	**Pool B**	D_{JS} (bits)
France	Tripolitania	0.5201
Libya	Croatia	0.5238
Slovenia	Fezzan	0.5295
Italy	Tripolitania	0.5308
Slovenia	Cyrenaica	0.5315
Croatia	Tripolitania	0.5349
Libya	Slovenia	0.5441
Slovenia	Tripolitania	0.5565

Remark: The Global Maximum Moved, and Moved Predictably

Volume II's twenty-one-pool matrix put whole-country Libya against Slovenia at its single most divergent pair, $D_{\mathrm{JS}} = 0.5442$ (Table **??**; the present twenty-one-pool recomputation reproduces this almost exactly, at 0.5441, the same 0.0001-bit-level agreement seen throughout this recomputation). Adding Libya's three provinces does not merely preserve this finding, it sharpens it: Tripolitania – the province Section 25.3 identified as having the *lowest* entropy of Libya's three, and therefore the pool furthest, of any Libyan pool, from high-entropy Balkan Slavic toponymy – replaces whole-country Libya at the top of the ranking, at $D_{\mathrm{JS}} = 0.5565$, 0.0124 bits above whole-country Libya's own figure against the same partner. This is the twenty-six-pool matrix Chapter 25, §25.6 anticipated: sub-national delimitation does not dissolve Libya's outlier status by revealing a more-typical province underneath the national aggregate, it isolates the specific province responsible for the largest share of that outlier signal. Six of the table's eight most divergent pairs involve either Libya or one of its three provinces, against three different Balkan or Western European partners (Slovenia, Croatia, Italy, France) – the same pattern Volume II's twenty-one-pool matrix showed for whole-country Libya alone (every one of its six most divergent pairs involved Libya; Table **??**), now visibly concentrated in one province rather than spread evenly across all three.

Remark: Mauritania's Sub-Pools Do Not Reproduce This Pattern

Neither Sahel nor Sahara appears in either half of Table 28.1. This is consistent with Chapter 26, §26.4's own finding that both Mauritanian sub-pools' nearest external neighbours are ordinary Maghreb countries at moderate divergence (0.11–0.13 bits) rather than extreme outliers: Mauritania's whole-country pool was never among Volume II's six most divergent pairs in the first place (Table **??**), so there was no concentrated outlier signal for a sub-national split to isolate. Libya and Mauritania, despite both being new Maghreb-adjacent additions to this chapter, behave oppositely under finer delimitation for exactly the reason each country's own chapter already gave: one was an outlier pool to begin with, the other was not.

28.3 Hierarchical Clustering on the Full Matrix

Average-linkage (UPGMA) hierarchical clustering, identical in method to Volume II, §**??**, was applied directly to the full 26×26 Jensen–Shannon divergence matrix, including the five subset-containment pairs excluded from Table 28.1 above (their presence or absence makes essentially no difference to the resulting tree, since UPGMA distances are driven by the full matrix rather than any single pair, but they are included here for consistency with Section 28.1's stated method). The cophenetic correlation coefficient is 0.8756, closely comparable to Volume II's 0.859 for the smaller twenty-one-pool tree and indicating a similarly faithful, though still imperfect, one-dimensional summary of the full 325-entry matrix.

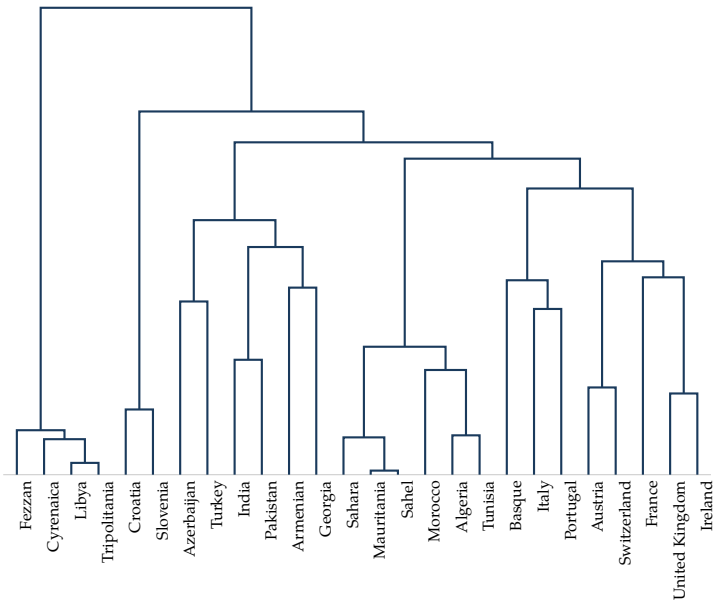

Figure 28.1: UPGMA dendrogram, twenty-six real pools, average-linkage clustering on the full 26×26 Jensen–Shannon divergence matrix (bits, vertical axis). Cophenetic correlation 0.8756. (Real data, this volume; `scipy.cluster.hierarchy`.)

Remark: Libya Forms Its Own Cluster, Separate From the Maghreb

Cutting Figure 28.1's tree at five clusters – the same cut Volume II, §**??** used – recovers: (i) Libya together with all three of its own provinces (Fezzan, Cyrenaica, Tripolitania), a tight, self-contained cluster distinct from every other pool in the matrix; (ii) Croatia and Slovenia; (iii) Azerbaijan, Turkey, India, Pakistan, Armenian, and Georgia; (iv) Mauritania together with Morocco, Algeria, Tunisia, and – notably – both of its own Sahel and Sahara sub-pools, the full Maghreb-plus-Sahel/Sahara group; and (v) the large mixed cluster of Basque, Italy, Portugal, Austria, Switzerland, France, United Kingdom, and Ireland. Two results follow directly from this partition, and both were only checkable once Libya's and Mauritania's sub-national pools were placed inside the same matrix as the whole-country ones. First, Libya's whole-country outlier status (Volume II, §**??**–§**??**) is not an artefact of pooling three internally very different provinces into one misleadingly uniform-looking national figure: all three provinces cluster *with each other*, and with whole-country Libya, before joining any other pool in the tree, meaning the outlier signal is a property shared broadly across Libya's constituent regions rather than concentrated in, or diluted by, any single one of them relative to the others. Second, Mauritania's Sahel and Sahara sub-pools cluster with Mauritania's own whole-country pool and with the rest of the Maghreb before joining anything else, directly confirming at sub-national resolution what Chapter 26, §26.4 already showed from each pool's own individual nearest-neighbour search: Mauritania's internal climatic split does not pull either half away from the broader Maghreb clustering documented since Volume II. The five real pools this chapter set out to place inside the matrix therefore behave in the two most different ways a sub-national split could: Libya's provinces reinforce and isolate an existing outlier signal into its own cluster, while Mauritania's regions are ab-

sorbed, without remainder, into an existing well-established one.

28.4 Reproducibility: What Was Actually Run

```
# Every one of the twenty-one whole-country distributions was
# recomputed directly, not copied, from the original dump files:
POOL_FILES = {
    'Libya': 'LY.txt', 'Croatia': 'HR.txt', 'Slovenia': 'SI.txt',
    'Austria': 'AT.txt', 'Switzerland': 'CH.txt', 'France': 'FR.txt',
    'United␣Kingdom': 'GB.txt', 'Ireland': 'IE.txt', 'Italy': 'IT.txt',
    'Portugal': 'PT.txt', 'Mauritania': 'MR.txt', 'Morocco': 'MA.txt',
    'Algeria': 'DZ.txt', 'Tunisia': 'TN.txt', 'Azerbaijan': 'AZ.txt',
    'Turkey': 'TR.txt', 'India': 'IN.txt', 'Pakistan': 'PK.txt',
    'Armenian': 'AM.txt', 'Georgia': 'GE.txt',
}
pools_p = {}
for name, fn in POOL_FILES.items():
    names = load_country_pool(fn)
    pools_p[name] = bigram_distribution(names)
pools_p['Basque'] = bigram_distribution(
    load_delimited_pool('ES.txt', {'PaisVasco': '59', 'Navarra': '32'}))

# The five new sub-national pools reuse Chapters 22-23's own verified
# admin1 code sets directly:
LIBYA_PROVINCES = {
    'Tripolitania': ['77','81','68','67','80','74','82','72','76'],
    'Cyrenaica':    ['69','66','63','70','79','83','65'],
    'Fezzan':       ['75','73','71','84','78','64'],
}
MAURITANIA_REGIONS = {
    'Sahel':  ['01','02','03','04','05','06','10'],
    'Sahara': ['07','08','09','11','12'],
}
for prov, codes in LIBYA_PROVINCES.items():
    pools_p[prov] = bigram_distribution(load_delimited_pool('LY.txt', codes))
for reg, codes in MAURITANIA_REGIONS.items():
    pools_p[reg] = bigram_distribution(load_delimited_pool('MR.txt', codes))

# All C(26,2) = 325 pairwise divergences, then UPGMA clustering exactly
# as Volume II, Appendix code_realtda21's cluster_pools function:
```

```
labels = list(pools_p.keys())
JS = np.zeros((26, 26))
for i, j in itertools.combinations(range(26), 2):
    JS[i, j] = JS[j, i] = js_divergence(pools_p[labels[i]], pools_p
       [labels[j]])
Z, coph_corr, dendro = cluster_pools(JS, labels)   # coph_corr =
   0.8756
```

Chapter 29

Conclusion: What Twenty-Six Pools Show That One Could Not

Methodological Transparency: What This Chapter Does, and Does Not, Do

Every figure cited below was already computed, verified, and discussed in the chapter it comes from; nothing here is a new computation. This chapter's only job is to step back from twenty-eight case-study and survey chapters and state, in one place, what recurred often enough across independent zones and independent methods to count as a volume-wide finding, what the volume's own methodology turned out to require correcting along the way, and what it has deliberately left for a fourth volume rather than pre-empted here.

29.1 The Preface's Aim, Revisited

This volume's Preface set a single, narrow aim: close Volume II's acknowledged gap between whole-country pools and genuine substrate zones, using only delimitation criteria verifiable inside the GeoNames files themselves. Twenty-six real, individually verified pools later – the twenty-one whole-country pools inherited from Volume II plus five sub-national zones (Libya's three historic provinces and Mauritania's Sahel/Sahara split) placed directly inside the same comparative matrix in Chapter 28 – that aim has been met at a scale the opening two-zone Anatolian test case did not attempt to predict. The method has not changed since Chapter 1: every zone in this volume was built from codes recovered from the data files' own `ADM1` or `ADM2` rows, or explicitly flagged as a declared geographic judgement (the Maghreb and Mauritania coastal/interior splits, Libya's tripartite historical assignment) rather than a delimitation the data themselves supply.

29.2 A Recurring Pattern: The Same Handful of Pools Keep Isolating Themselves

No single chapter set out to test this, but the same result surfaced independently, under different measures, in different parts of the volume: a small set of pools is repeatedly the most distant or the least clustered entry in whatever table or den-

drogram it appears in. Basque was identified as the least like every other pool in Chapter 14's cross-comparison (§14.3); Sindh stood apart from every other pool in Chapter 16's table (§16.3); Corsica's isolation was noted, in Chapter 17, as mirroring the Basque and Sindh pattern directly (§17.3); Sardinia was the most distant pool from every other pool in Chapter 19's table (§19.3); and Libya, at whole-country resolution since Volume II and confirmed at sub-national resolution in Chapter 28's dendrogram (§28.3), forms its own cluster before joining any other pool in the tree. These are five different pools, delimited in five different chapters, compared under at least three different measures (Jensen–Shannon divergence, Fréchet-mean Fisher–Rao distance, and UPGMA clustering on the full matrix). None of this volume's chapters claims this as a single unified phenomenon – Basque, Sindh, Corsica, and Sardinia are conventionally understood as linguistic or historical isolates for reasons external to this series' data, while Libya's isolation was traced, in Chapter 28, to a signal shared broadly across its own three provinces rather than to any one of them – but the fact that independently delimited, independently computed pools keep landing in the same structural role across chapters is itself a volume-wide observation this Conclusion is the first place to state plainly.

29.3 A Recurring Caution: Two Measures Do Not Always Agree, and Neither Is Simply Right

A second pattern recurred often enough to be a methodological lesson rather than a one-off finding. Chapter 15 recorded a genuine, unresolved disagreement between Jensen–Shannon divergence and Fréchet-mean Fisher–Rao distance over which Balkan zone was closer to Croatia (§15.3); Chapter 19 found the same two measures naming two different regions as Italy's "most distinctive" (§19.4); and Chapter 21 found that the Fréchet-distance figures alone missed a point the Jensen–Shannon ranking made visible (§21.2). This volume's stated response, every time, was the same: report both measures, record the disagreement as genuine rather than adjudicate in favour of either, and treat it as a predicted consequence of Section 14.4's methodological lesson rather than a defect in one measure. The Jensen–Shannon survey chapters went further and uncovered an actual methodological confound of this kind: Section 14.4 first documented that a delimited pool's divergence against its own enclosing whole-country baseline is mechanically depressed the larger a share of that baseline the pool occupies, and Chapter 22 (§22.3) showed the same pattern holds exactly, and without exception, across all three Maghreb countries' coastal/interior splits – a correction this volume made openly, keeping the original published pool sizes where they were already correct (§22.2) rather than discarding them.

29.4 Elevation: A Genuine but Narrow Data-Native Criterion

Chapter 5 introduced elevation as a second data-native delimitation criterion alongside `admin1`/`admin2` codes, and the volume's later chapters tested it wherever the data allowed. The verdict is mixed rather than uniformly positive, which is itself

worth stating plainly: elevation moved name length but not entropy in its first test (§5.3); the sign of its effect on entropy flipped, and was traced to a specific, narrow threshold, in the Caucasus chapters (§11.3); and it proved inapplicable to Libya outright, where the `elevation` field covers under 1.2% of the national pool (§25.7). Elevation earned its place as a legitimate second criterion in this series precisely because the volume reports these mixed and negative results alongside the positive ones, rather than only the chapters where it worked cleanly.

29.5 What Is Deliberately Left Open

Every chapter's own Roadmap section already states its own unresolved items in full, and this Conclusion does not repeat them individually. Three items recur across more than one chapter's roadmap and are therefore the most natural starting points for a fourth volume: finer `admin2`-level splits of zones this volume only delimited at `admin1` resolution (flagged separately for Trentino-Alto Adige in Chapter 19, for historic Occitania in Chapter 17, and for Croatian Istria in Chapter 15); resampling checks on small-n pools whose distinctiveness has not yet been separated from sampling noise (Islamabad in Chapter 16; Nouakchott in Chapter 26); and extension of this volume's already-delimited-country chapters to their remaining `admin1` units not yet tested here, of which Chapter 18's twenty-six untested Indian states is the largest single instance. None of these is attempted in this volume, for the same reason no chapter in it invents a result it has not verified: each requires either new subnational delimitation work or a resampling protocol not yet run, and this volume's closing Appendix A exists precisely so that a fourth volume, or anyone else, can rerun the same verified pipeline on them rather than take this Conclusion's word for what such a delimitation would show.

Appendix A

Reproducibility: The Zone-Verification and Delimitation Code

```
# 1. Recover a province's own admin1 code from its self-reported
#    ADM1-feature row -- the verification step of Section 1.1,
#    required before any admin1-based delimitation in this volume.
def find_admin1_code(path, province_name):
    with open(path, encoding='utf-8') as f:
        for line in f:
            parts = line.rstrip('\n').split('\t')
            if len(parts) < 11:
                continue
            name, fclass, fcode, admin1 = (parts[2], parts[6],
                                           parts[7], parts[10])
            if fclass == 'A' and fcode == 'ADM1' and name ==
               province_name:
                return admin1
    return None   # not found: do NOT guess: the zone cannot be
       built

# 2. Build the verified code set for a named zone
def build_zone_codes(path, province_names):
    codes = {}
    for prov in province_names:
        code = find_admin1_code(path, prov)
        if code is None:
            raise ValueError(f"admin1 code for {prov!r} not found
               in "
                             f"{path}; cannot delimit this zone
                                from "
                             f"the data at hand")
        codes[prov] = code
    return codes

```

```
# 3. Restrict the pool to the verified admin1 code set (feature classes
#    P and T, normalised name length >= 3 -- unchanged from Volume II)
def load_delimited_pool(path, admin1_codes):
    codes = set(admin1_codes.values())
    names = []
    with open(path, encoding='utf-8') as f:
        for line in f:
            parts = line.rstrip('\n').split('\t')
            if len(parts) < 11:
                continue
            if parts[6] not in ('P', 'T'):
                continue
            if parts[10] not in codes:
                continue
            az = normalize(parts[2])          # Volume II, Appendix 8.1
            if len(az) >= 3:
                names.append(az)
    return names

# 4. Everything downstream (aggregate entropy, n=40 Fisher-Rao sample,
#    Frechet mean, Vietoris-Rips persistence) is Volume II, Appendix 8's
#    unchanged pipeline, applied to `load_delimited_pool`'s output
#    exactly as it was applied to a whole-country pool in Chapter 9.
```

Remark: The One New Idea in This Appendix

Every routine here is either identical to Volume II's or a thin wrapper around it, with a single genuine addition: `find_admin1_code`'s refusal to return a guessed value. `build_zone_codes` raises an exception rather than silently falling back to an assumed code if a province name is not found as its own `ADM1` row – a deliberate design choice matching this volume's stated methodological commitment that a zone which cannot be verified from the data at hand should not be built from it, rather than approximated.